LANL report: LA-UR-95-1265

NUCLEAR SCIENCE AND ENGINEERING: **127**, 54–77 (1997)

Confidence Interval Procedures for Monte Carlo Transport Simulations

S. P. Pederson

Georgia Institute of Technology, School of Industrial and Systems Engineering
Atlanta, Georgia 30332-0205

and

R. A. Forster and T. E. Booth

Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Received April 20, 1995
Accepted February 24, 1997

Abstract—*The problem of obtaining valid confidence intervals based on estimates from sampled distributions using Monte Carlo particle transport simulation codes such as MCNP is examined. Such intervals can cover the true parameter of interest at a lower than nominal rate if the sampled distribution is extremely right-skewed by large tallies. Modifications to the standard theory of confidence intervals are discussed and compared with some existing heuristics, including batched means normality tests. Two new types of diagnostics are introduced to assess whether the conditions of central limit theorem–type results are satisfied: The relative variance of the variance determines whether the sample size is sufficiently large, and estimators of the slope of the right tail of the distribution are used to indicate the number of moments that exist. A simulation study is conducted to quantify the relationship between various diagnostics and coverage rates and to find sample-based quantities useful in indicating when intervals are expected to be valid. Simulated tally distributions are chosen to emulate behavior seen in difficult particle transport problems. Measures of variation in the sample variance s^2 are found to be much more effective than existing methods in predicting when coverage will be near nominal rates. Batched means tests are found to be overly conservative in this regard. A simple but pathological MCNP problem is presented as an example of "false" convergence using existing heuristics. The new methods readily detect the false convergence and show that the results of the problem, which are a factor of 4 too small, should not be used. Recommendations are made for applying these techniques in practice, using the statistical output currently produced by MCNP.*

I. INTRODUCTION

Monte Carlo simulations performed on radiation transport codes such as MCNP (Ref. 1) are extensively used in nuclear engineering to model the behavior of systems of particles and to obtain estimates of certain physical parameters, such as the mean fraction of particles entering or leaving a given region. These simulations typically employ variance reduction schemes that result in sampled tally distributions that can be extremely right-skewed. This skewness can make estimation and confidence interval formation of underlying population parameters difficult, as it may take a long time to obtain a sample representative of the entire underlying distribution. In this paper we examine the problem of obtaining confidence intervals for means under these conditions.

To illustrate, consider the following test problem originally discussed in Forster, Booth, and Pederson.[2] The quantity of interest was the surface neutron leakage flux above 12 MeV from an isotropic 14-MeV neutron point source of unit strength at the center of a 30-cm-thick concrete shell with an outer radius of 390 cm. The variance reduction techniques used were implicit capture with weight cutoff, low-score point detector Russian roulette, and a 0.5-mean-free-path (4-cm) neighborhood around a point detector. The correct answer, based on a long (1-billion-history) run with the more stable ring detector, is 5.64×10^{-8} n/cm^2·s ± 0.02%. Table I shows a tally fluctuation chart (TFC) from point detector tally 35 for a run of 20 000 independent sample histories (i.e., $n = 20\,000$); a TFC tracks various sample statistics over time. Figure 1 shows a log-log probability density plot of the data. The solid line represents the 1-billion-history point detector density, the long-dashed line the 1-billion-history ring detector density, and the short-dashed line the point detector density after 14 000 histories. Both the TFC and density plot are statistical tools that are currently available in MCNP, Version 4A.

The result for $n = 1000$ histories has a small relative error (RE) of 0.73%. An experienced Monte Carlo practitioner might well have accepted the result at 14 000 histories as valid, with small RE and fairly constant figure of merit (FOM). However, jumps in statistics between 14 000 and 15 000 and again between 19 000 and 20 000 indicate that the estimation has not stabilized by this time. Not only does the sample mean appear unstable (it jumps by nearly 25% as n goes from 19 000 to 20 000), it grossly underestimates the true mean (by a factor of 4). Confidence intervals for the mean flux fall short of the true value by a factor of 3. All the statistics shown should be stable and relatively unbiased if the problem has converged. Furthermore, the density plots in Fig. 1 indicate that there are areas of the tally distribution, basically any scores $>10^{-5}$, that have not been adequately sampled with 14 000 histories. Should this result be accepted? The new diagnostics discussed herein, and alluded to in this example, will clearly indicate that sampling needs to continue, as estimation of the flux has not yet stabilized.

TABLE I

Tally Fluctuation Chart for Example Problem, from Point Detector Tally 35
[$\mu = 5.64 \times 10^{-8}(\pm 0.02\%)$]

n	$\bar{x}$	RE	VOV	FOM	SLOPE
1 000	1.2806E−08[a]	0.0073	0.9838	122 708	1.8
2 000	1.3779E−08	0.0623	0.9332	810	1.7
3 000	1.3576E−08	0.0432	0.8530	1 096	1.7
4 000	1.3496E−08	0.0342	0.7144	1 289	1.7
5 000	1.3897E−08	0.0415	0.4327	694	1.6
6 000	1.3701E−08	0.0351	0.4328	803	1.5
7 000	1.4625E−08	0.0730	0.7264	157	1.5
8 000	1.4395E−08	0.0649	0.7264	174	1.5
9 000	1.4237E−08	0.0584	0.7247	191	1.4
10 000	1.4133E−08	0.0530	0.7186	208	1.4
11 000	1.4017E−08	0.0486	0.7183	224	1.4
12 000	1.4004E−08	0.0451	0.6869	238	1.4
13 000	1.4212E−08	0.0437	0.5425	234	1.4
14 000	1.4124E−08	0.0409	0.5421	249	1.4
15 000	1.5084E−08	0.0781	0.6495	63	1.3
16 000	1.4993E−08	0.0737	0.6479	67	1.3
17 000	1.4872E−08	0.0700	0.6479	70	1.3
18 000	1.4849E−08	0.0663	0.6423	73	1.3
19 000	1.4805E−08	0.0631	0.6388	77	1.3
20 000	1.8390E−08	0.1464	0.4065	14	1.3

[a] Read as 1.2806×10^{-8}.

In problems such as this, the relevant question to ask is whether the user has run the code for a sufficiently long period of time to ensure reasonably valid inference about the parameter of interest. In statistical terms, this is equivalent to asking if the sample size n (i.e., number of histories) is sufficiently large so that the standardized sample mean follows an approximately normal (Gaussian) distribution, and thus confidence intervals for the parameter, using the first two sample moments, are valid.

This paper proceeds as follows: Sec. II discusses interval estimation for means and describes some heuristics used in MCNP4A to indicate when enough data have been obtained. Improvements to these methods, based on higher order moment expansions and characterizations of the tails of distributions, will be examined. We also consider other methods currently used, specifically, batched means and normality tests. Section III describes a simulation study investigating these new methods and some techniques for predicting coverage rate validity. The paper concludes with some recommendations for the Monte Carlo transport code user and the application of these recommendations to the example problem.

While the use of higher moment estimators to assess the convergence of means in non-Gaussian settings has existed for many years (see Ref. 3 for an early treatment), the use of tail characterization parameters (such as the slope) is a more recent development (e.g., Ref. 4). Much of this work grew out of the problem of predicting excedances over thresholds (e.g., the likelihood of water levels cresting over flood stage). There is now a great deal of interest in extremely long-tailed distributions in fields such as telecommunications and finance (Ref. 5 models waiting times for Ethernet traffic), but up to now the primary interest has been in the characterization of the distribution and not in the properties of the sample mean (other than the question of whether the underlying population mean is finite). We believe this paper is the first attempt to use these two methodologies in combination to obtain confidence intervals for means of severely skewed distributions.

This paper focuses on highly nonnormal underlying distributions. None of the methods we propose hamper estimation in well-behaved cases (i.e., the skewness correction factor converges to zero as the underlying distribution becomes more symmetric). Rather, our methods attempt to improve and characterize the convergence to approximate normality of sample means from highly

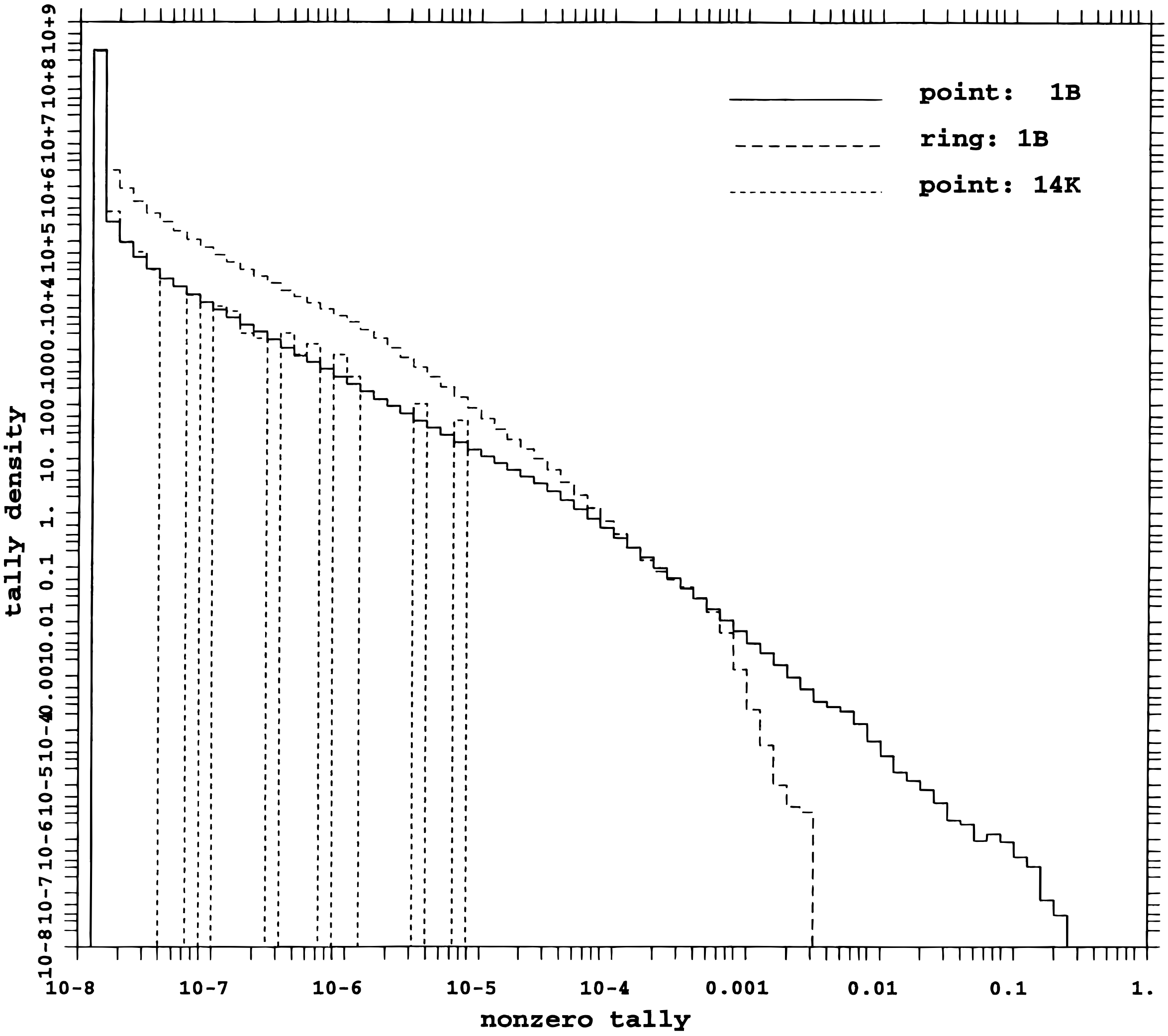

Fig. 1. Tally 35 histogram.

non-Gaussian distributions. Obviously, because these methods are data driven, if a portion of the phase-space is unsampled, our techniques will not account for the corresponding contribution to the quantity of interest μ. However, no data-driven method will do so. There is always the chance of a "silver bullet" region of probability that has not been sampled, but the likelihood of such an event cannot be determined without auxiliary information.

II. CURRENT PRACTICE AND NEW RESULTS

Monte Carlo particle transport codes are invaluable tools. They simulate processes that are too costly or difficult to physically measure or that have no direct analytical solution. As seen in the example, a code is typically run for a predetermined number of histories n or computer time T, and tallies are kept for the estimates of interest. These calculated tallies are sample means that estimate some underlying physical parameter μ. In what follows, we will model the tally quantity of interest as a random variable X with distribution function $F(x) = P(X \leq x)$ and density function $f(x) = F'(x)$, the latter being defined where $F(x)$ is continuous. The probability of X falling in the small interval $(x, x + dx)$ is $f(x)\,dx$. When there are discontinuities in F, corresponding to probability at a point as opposed to an interval, f is augmented by the probabilities of those discrete points, i.e., by $p(x) = P(X = x)$, so that the following definition holds: $\int f(x)\,dx = 1$. The expectation or mean of X is $\mu = \int x dF(x) = \int x f(x)\,dx$. We will denote the sample mean based on n independent and identically distribution (iid) observations or histories $x_1, \ldots, x_n$ as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ , \tag{1}$$

which is usually taken as the point estimate of μ. The variance σ^2 [$= \int (x - \mu)^2 \, dF(x)$] of X is estimated by the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \approx \frac{1}{n} \sum x_i^2 - \bar{x}^2 \quad , \tag{2}$$

where the last formula applies for large n. For random variables of only one sign (either positive or negative), the relative error, or coefficient of variation, is used to measure the variation in X relative to its mean and is simply the ratio $s/\bar{x}$. The analogous relative error for the sample mean $\bar{x}$ is given by

$$\mathrm{RE} = \frac{s}{\sqrt{n}\bar{x}} \quad , \tag{3}$$

where $s/\sqrt{n}$ is the standard error of $\bar{x}$. An associated quantity, the FOM, is defined as

$$\mathrm{FOM} = \frac{1}{\mathrm{RE}^2 \times T} \quad , \tag{4}$$

where T is the total computer time taken to obtain n histories. For the distributions considered here, T will be directly proportional to n. In this case, the FOM is proportional to the squared signal-to-noise ratio; i.e., FOM $\propto \bar{x}^2/s^2$. Signal-to-noise ratios represent the amount of information available about μ, and in general, the larger the signal-to-noise ratio, the better.

II.A. Confidence Intervals

Confidence intervals for μ are formed by assuming that $\bar{x}$ has an approximately normal (Gaussian) distribution for large samples. Specifically, for a sequence of iid observations $x_1, \ldots, x_n$, and assuming that at least two moments of X are finite, the central limit theorem states that

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim \mathcal{N}(0,1) \tag{5}$$

as $n \to \infty$, where $\mathcal{N}(0,1)$ denotes a standard normal random variable; i.e., it has zero mean and unit variance. In general, σ^2 is not known, and in practice it is estimated by s^2. It can also be shown, with the same assumptions, that the pivotal statistic

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \tag{6}$$

has a limiting $\mathcal{N}(0,1)$ distribution. However, unless four moments of the underlying distribution are finite, the rate of convergence of t to normality can be much slower than $n^{-1/2}$, because s^2 is not guaranteed to converge to σ^2 at rate n^{-1} (see Refs. 6 and 7). Inverting t obtains 100 × $(1 - \alpha)$% two-sided confidence intervals $\mathcal{I}$ for μ of the form

$$\mathcal{I} = (l_\alpha, u_\alpha) = (\bar{x} + z_{\alpha/2} s/\sqrt{n}, \bar{x}_n + z_{1-\alpha/2} s/\sqrt{n}) \quad , \tag{7}$$

where z_ν represents the $100 \times \nu$ percentage point of the standard normal distribution. One-sided intervals are computed similarly; the upper $100 \times (1 - \alpha)$% interval for μ is

$$\mathcal{I}_u = (-\infty, \bar{x} + z_{1-\alpha} s/\sqrt{n}) \quad , \tag{8}$$

where the left end point can be replaced by zero if X is nonnegative.

When severe skewness is present in the data, these asymptotic approximations may be inaccurate in what are commonly thought to be "large-sample" situations (see Ref. 8, for example). Confidence intervals are said to be valid if they cover the parameter of interest the nominal (i.e., $1 - \alpha$) fraction of the time.[9] Upper limits tend to cover at rates lower than $1 - \alpha$ if right-skewness exists, as all moment estimates are biased downward (i.e., they tend to underestimate the corresponding population moment), and hence the upper end point of $\mathcal{I}$ or $\mathcal{I}_u$ is too small on average. Lower limits are less sensitive because the biases in $\bar{x}$ and s partially cancel out. Another commonly used measure of confidence interval performance is the expected half-width of the interval (see Ref. 10). The expected half-width is estimated by $z_{1-\alpha/2} s/\sqrt{n}$, which is negatively biased (too small) in the presence of skewness.

Low coverage rates are generally the result of too few large x_i being observed, not too many; in the example from Sec. I, the estimates and confidence intervals were three times too low, as not enough large observations had been sampled, even for $n = 20\,000$. This perhaps runs counter to intuition, as it is often the jumps in statistics monitored over time that cause the concern that the underlying process is being inadequately sampled. Extremely large scores x_i will occur from time to time and may shift the sample mean dramatically. However, the impact of these large observations should lessen as n becomes large. These occasional shifts are in fact the necessary rare contributions from the tail of the distribution. What is also necessary for valid inference is some indication that the conditions of the large-sample theory apply, in particular that at least two and preferably four moments are finite. If the variance is infinite, confidence intervals will not have the behavior prescribed by the central limit theorem, even for large n. In this case, the sample mean $\bar{x}$ will converge to μ, but at an indeterminate rate. Intervals will grow large asymptotically, resulting in a coverage probability of 1, but coverage rates are not guaranteed to be conservative for any finite sample size.

Heuristic rules have evolved over time to give the user guidance concerning when the large-sample normal approximation is valid; see Forster, Pederson, and Booth.[11] The most popular of these relies on the relative error of $\bar{x}$ and states that an RE of 0.10 (0.05 for a point detector tally) is the largest value consistent with a converged solution. This result is based on the fact that if a two-point (say 0 and 1) distribution were considered, one hundred 1's would correspond to an RE of 0.10; this number of 1's provides approximate normality for t. However, for an arbitrary distribution, a similar heuristic is less obvious. [Note also that there is a direct relationship between the half-width of a confidence interval and RE. In particular, a $100 \times (1-\alpha)\%$ interval has relative half-width $z_{1-\alpha/2}$RE.]

II.B. Other Methods

II.B.1. Mixture Distributions

Lux and Koblinger[12] describe a method for efficiently estimating the sample mean of a distribution that is the mixture of point mass at zero and a (nonzero) Gaussian distribution. While many relevant distributions in neutron particle transport (especially those employing variance reduction techniques) can be considered as a similar sort of mixture distribution, the Gaussian assumption is untenable for the distributions we are considering. In particular, if the nonzero score tail is Gaussian, then the mixture distribution of zero and nonzero scores behaves essentially like a binomial distribution. A binomial distribution is very well behaved, and its corresponding sample mean quickly converges to a Gaussian. That is, standard confidence interval estimation methods work well.

II.B.2. Batched Means and Normality Tests

The methods of batched means and normality tests are often used to assess the convergence of a mean to normality. The idea behind each is simple. The method of batched means replaces the original sequence of data $x_1, x_2, \ldots,$ with a sequence of batched means $\bar{x}_1^*, \bar{x}_2^*, \ldots,$ where each $\bar{x}_i^*$ is the mean of B successive observations. This sequence of batched means is now treated as the sequence of data, and variance estimation is based on the new sample of size $m = n/B$ (Refs. 12 and 13). Generally, both m and B grow as n does. This method is especially useful when the data stream is correlated, as it improves estimation of the variance of the sample mean. However, when the data are independent, as in the cases discussed in this paper, Refs. 12 and 14 give derivations showing that the optimal batch size for variance estimation is $B = 1$. That is, in independent sampling situations, batching means is less efficient than not batching.

Another use of batched means is in the determination of normality of the distribution of the overall sample mean. The idea behind normality tests is that if a sample does not appear to be statistically significantly different from a Gaussian distribution, one may assume that normality is an adequate representation of the distribution in question. Two commonly used normality tests are the Wilk-Shapiro test (a formal hypothesis test) and the use of q-q (quantile-quantile) normal plots.[15] The Wilk-Shapiro test and q-q plots each compare the ordered observations of a sample with the expected quantiles of a Gaussian distribution. A modified version of the Wilk-Shapiro statistic that is much easier to compute is[15]

$$W' = \frac{(\boldsymbol{u}^T\boldsymbol{a})^2}{(\boldsymbol{a}^T\boldsymbol{a})\sum (u_i - \bar{u})^2} \ , \qquad (9)$$

where $\boldsymbol{u}$ is a vector of data $\{u_i\}$ of length r with mean $\bar{u}$, and $\boldsymbol{a}$ is a vector of expected order statistics from a standard Gaussian distribution. That is, a_i is the expected value of the i'th ordered observation from a sample (of size r) from a standard Gaussian. W' is the squared correlation between $\boldsymbol{u}$ and $\boldsymbol{a}$ and will be near 1 (1 = perfect correlation) for a sample that is nearly normal. Many other goodness-of-fit tests are available; see Ref. 12. Q-q plotting is a graphical technique similar in spirit to the Wilk-Shapiro test, in that one plots the ordered observations from the sample against the ordered quantiles from a normal distribution. If no large departures from normality are present, this plot should approximate a straight line.

Returning now to the problem of whether the mean arising from a highly skewed distribution has an approximately Gaussian distribution, a common technique in detecting nonnormality is to use one of the normality tests on a set of batched means. In this case, inferences are not on either the original distribution or the distribution of $\bar{x}$ but rather on the distribution of batched means $\bar{x}_1^*, \bar{x}_2^*, \ldots,$ each of size B. Thus, if a batched mean test indicates normality, one can infer that the overall mean, based on a larger sample, should also follow a Gaussian distribution. However, the converse is not necessarily true. If the batched mean test rejects normality, there is no definite indication regarding the convergence of $\bar{x}$. In the situations described in this paper, we deal with extremely skewed distributions, so the overall mean will converge much sooner than the batched means do. Hence batched means tests are not an efficient method for detecting normality of overall means generated in highly skewed situations. Section III.C.1 examines a simulation of W' and discusses its utility as an indicator of normality of the overall sample mean $\bar{x}$.

The remainder of this section will discuss three types of improvements to the methods presented so far. To begin, higher moment quantities such as the relative variance of the variance (VOV) are used to characterize the convergence of t to normality. Next, higher moment modifications to the standard confidence intervals are discussed. The section closes with methods for incorporating information about the shape of the tail of the density $f(x)$.

II.C. Higher Order Approximations

To improve the performance of confidence intervals when skewness is present, additional moment information is needed. We will consider expansions for t based on higher order moments (when they exist) and also models for characterizing the right-hand tail of the density $f(x)$ of X.

Edgeworth expansions[16] are power series representations of distributions of sums of iid random variables in terms of higher order moments. They are useful in characterizing departures from normality. Expansions are also available for functions of these sums, such as the pivotal statistic t, a function of $\sum x_i$ and $\sum x_i^2$. The cumulative distribution function (cdf) of t, $G_t(x)$, can be written as

$$\begin{aligned} G_t(x) &= P(t \leq x) \\ &= \Phi(x) + \phi(x)[n^{-1/2}q_1(x) + \cdots + n^{-j/2}q_j(x)] \\ &\quad + o(n^{-j/2}) \ , \end{aligned} \tag{10}$$

where $\Phi(x)$ and $\phi(x)$ are the distribution and density functions, respectively, for a standard normal, and $o(u)$ represents a term for which $o(u)/u \to 0$ as $u \to 0$. This expansion in j terms is valid if $j + 2$ moments exist and the distribution function $F(x)$ is absolutely continuous (i.e., X has a continuous density function). When these conditions do not exist, it is not possible to characterize the remainder in this manner (i.e., terms of all orders of $n^{-1/2}$ exist in the remainder), and the rate of convergence to normality is indeterminate. The functions q_j are polynomials whose coefficients are functions of the moments of X. The first two q_j are

$$q_1(x) = \frac{1}{6}\frac{\mu_3}{\sigma^3}(2x^2 + 1) \tag{11}$$

and

$$q_2(x) = x\left[\frac{1}{12}\frac{\mu_4}{\sigma^4}(x^2 - 3) - \frac{1}{18}\frac{\mu_3^2}{\sigma^6}(x^4 + 2x^3 - 3) - \frac{1}{4}(x^2 + 3)\right] \ , \tag{12}$$

where $\mu_3\,[=\int(x-\mu)^3\,dF(x)]$ and $\mu_4\,[=\int(x-\mu)^4\,dF(x)]$ are the third and fourth central moments, respectively, defined analogously to $\mu_2 = \sigma^2$. From these formulas we can see the direct dependence of the coverage rate on the scaled higher order moments. Formulas for estimates of μ_3 and μ_4 are direct analogs of the sample variance formula (n, as opposed to $n - 1$, is used as the divisor for large n):

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n}\left(\sum x_i^3 - 3\bar{x}\sum x_i^3\right) + 2\bar{x}^3 \ , \tag{13}$$

$$\begin{aligned} \hat{\mu}_4 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \\ &= \frac{1}{n}\left(\sum x_i^4 - 4\bar{x}\sum x_i^3 + 6\bar{x}^2\sum x_i^2\right) - 3\bar{x}^4 \ . \end{aligned} \tag{14}$$

Alternatively, we can obtain moment approximations for t via application of the delta method.[16] First, we will reparameterize in terms of moments of $\bar{x}$ and s^2. In particular, define the correlation between $\bar{x}$ and s^2 as

$$\rho^2 = cor(\bar{x}, s^2) = \frac{\mu_3^2}{\sigma^2(\mu_4 - \sigma^4)} \tag{15}$$

and the squared coefficient of variation of s^2 as

$$\gamma^2 = cv^2(s^2) = \frac{\mu_4 - \sigma^4}{n\sigma^4} \ . \tag{16}$$

From Ref. 17, the mean and next three central moments of t are

$$E(t) = -\frac{\rho\gamma}{2} + O(n^{-3/2}) \ , \tag{17}$$

$$\begin{aligned} Var(t) &= E[t - E(t)]^2 \\ &= 1 + \frac{3}{n} + \frac{7}{4}\rho^2\gamma^2 + O(n^{-2}) \ , \end{aligned} \tag{18}$$

$$E[t - E(t)]^3 = -2\rho\gamma + O(n^{-3/2}) \ , \tag{19}$$

and

$$E[t - E(t)]^4 = 3 + \frac{6}{n} - 2\gamma + 12\rho^2\gamma^2 + O(n^{-2}) \ , \tag{20}$$

where $O(u)$ represents a term for which $O(u)/u$ is bounded as $u \to 0$. This parameterization is useful when skewness is present. Solving Eqs. (15) and (16) for $\mu_3/n\sigma^3$ gives

$$\frac{\mu_3}{n\sigma^3} = \rho\gamma \leq \gamma \ , \quad 0 < \rho \leq 1 \ . \tag{21}$$

With high skewness, ρ, the correlation between $\bar{x}$ and s^2, is close to 1, and the bound in Eq. (21) is quite tight. This suggests that in addition to the scaled third moment, the coefficient of variation of s^2 can be used to characterize departures from normality of t. Sargent, Kang, and Goldsman[10] found empirical evidence that the variability of s^2 is useful in this regard.

These results and empirical work[18] have led Monte Carlo researchers to employ the VOV (Ref. 19), or the estimated relative variance of s^2,

$$\text{VOV} = \hat{\gamma}^2 = \frac{\hat{\mu}_4 - s^4}{s^4 n} \ . \tag{22}$$

(See Ref. 20 for discussion of a related term, the figure of reliability, which substitutes VOV for RE in the FOM formula.) Small values of VOV should correspond to t having mean and variance near 0 and 1, respectively, with low skewness. Some limited simulation results in Pederson[17] found that a VOV < 0.1 corresponds to near-nominal upper interval coverages at $\alpha = 0.05$. References 21 and 22 cite similar heuristics involving the scaled third moment directly, stating that $\gamma\rho$ should be much less than one for the normality result to apply. Which measure is used matters little when the correlation between $\bar{x}$ and s^2 is near 1, and we will continue to use VOV in what follows because of its additional interpretation as the relative variance of s^2. Even in skewed situations with ρ near 1, the variability of s^2 is the determining factor in the convergence of t to normality. It should be noted again that third- and fourth-moment estimators such as VOV are also biased low because of the skewness in $f(x)$, and hence they converge more slowly to their expectations than does $\bar{x}$ or s^2.

Illustrating the relationship between the first two sample moments of long-right-tailed distributions, Fig. 2 is a plot of 500 sample standard deviations versus sample means, generated under a right-skewed distribution (Pareto) for $n = 8000$ with a mean of 5.4 (indicated by a dashed vertical line), standard deviation of 55, and $\rho = 0.8$. For a symmetric underlying distribution with this mean and standard deviation, this plot should look roughly like an ellipse with axes parallel to the x and y axes, with $\bar{x}$'s centered at 5.4 and s's centered at 55 and themselves having standard error of ~1 (as opposed to the observed standard error of 18). A contour of equal probability from such a distribution is superimposed on the plot. Any points lying inside the V drawn on the plot correspond to confidence intervals (for a given α level, in this case 0.10) that cover the true mean μ; $\sim\alpha/2 = 0.05$ of the points should fall outside each arm of the V. In reality, 16% of the points fall to the left of the V, and 2% fall to the right. (There are three representative confidence intervals included on the plot, one of which covers the true mean, one of which misses on the low side, and one of which misses on the high side.) The points falling to the left of the V thus have upper intervals with an observed noncoverage rate of more than three times the nominal rate. The tilt of the data cloud corresponds to a nonzero correlation ρ that persists as n increases. The variation (and asymmetry) in s is also clearly seen. However, this variation drops as n increases, and as the s values cluster around their expected value (55 in this case), the fraction of points lying on each side of the V approaches nominal limits. The point here is that while the correlation between $\bar{x}$ and s^2 is present whenever sampled data are skewed, it is the additional variability in s^2 that distorts confidence interval coverages.

II.D. Modified Confidence Intervals

Several alternative confidence interval procedures have been proposed for high-skewness cases. Johnson[23] and Hall[24] each derive modified intervals based on a Cornish-Fisher expansion of t. Hall[25] presented two alternatives based on a monotonic transformation of an Edgeworth expansion of t; these transforms have coverage rate error $O(n^{-1})$, whereas the Cornish-Fisher–based expansions have error $O(n^{-1/2})$. The upper interval from Ref. 25 we consider is

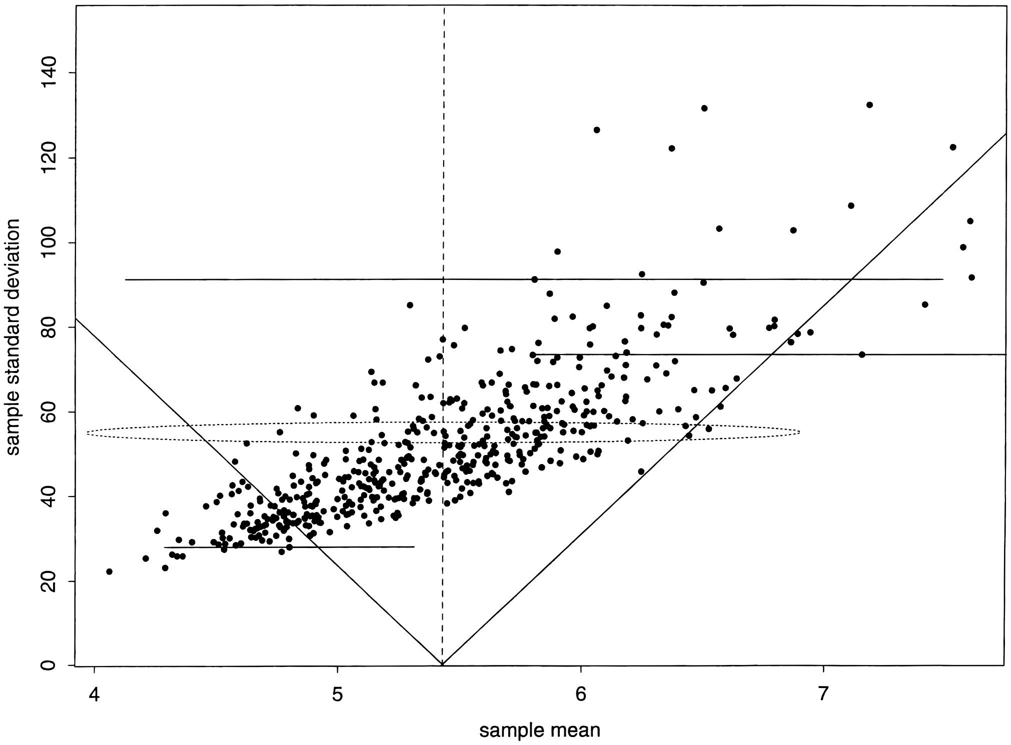

Fig. 2. Nominal 90% intervals for 500 Pareto samples, $n = 8000$, $\rho = 0.80$.

$$\mathcal{I}_u^* = \left(-\infty, \bar{x} - \frac{3\hat{\sigma}}{\hat{\zeta}}\left\{\left[1 + \hat{\zeta}\left(\frac{z_\alpha}{\sqrt{n}} - \frac{\hat{\zeta}}{6n}\right)\right]^{1/3} - 1\right\}\right) , \tag{23}$$

where $\hat{\zeta} = \hat{\mu}_3/\hat{\sigma}^3$ is the scaled third-moment estimate. This interval approaches $\mathcal{I}_u$ as $\hat{\zeta} \to 0$ for fixed n and is valid under the usual assumptions, the existence of at least three moments, and the existence of a proper density function (equivalently, a continuous cdf) for X.

II.E. Tail Characterization

The foregoing discussion assumes that sufficient (usually three or four) moments exist. In some Monte Carlo simulations this is not necessarily true, except for μ (and generally σ^2, but this is not guaranteed). One way of determining how many moments exist is to model the convergence of the sample moments (e.g., $\hat{\mu}_3$ or $\hat{\mu}_4$). Alternatively, directly modeling the tail of $f(x)$ can provide information about the existence of moments. We consider distributions that follow a power law, i.e., $P(X > x) \propto x^{-\kappa+1}$, for $x > x_0$, where x_0 is a threshold. The parameter κ indexes how many moments of the distribution are finite: $\kappa > k$ implies that $k - 1$ moments exist. This can be shown by noting that for power-law distributions, $f(x) \propto x^{-\kappa}$, and thus $E(X^k) \propto \int_{x_0}^{\infty} x^{k-\kappa}\,dx$, which is infinite for $k \geq \kappa - 1$.

A useful graphical tool for displaying the information in the tail of a distribution is a log-log histogram (e.g., Fig. 1), also called a log-log empirical density plot (of X). Histograms are bar plots with the area of the bars equal to the relative frequency of data in the corresponding x interval; for equally spaced intervals, the relative frequency is equal to the height of the bar. As the sample size and number of intervals become large, the histogram converges to the true density $f(x)$. Log-log histograms are just regular histograms (or density plots) with log scales for both axes. For the density of a power law with parameter $\kappa - 1$, the tail plotted in log-log scale will be a straight line with slope $-\kappa$; in what follows, we will use κ to denote what we refer to as slope (and will use $\hat{\kappa}$ or SLOPE to denote the corresponding estimator). The code MCNP4A supplies log-log histograms and estimated slopes of the sampled distributions, useful diagnostics for detecting whether the tail slope appears to drop off sufficiently quickly (e.g., SLOPE > 5 implies that at least four moments exist, if a power law is appropriate). Log-log histograms can also show if change points occur in the tail slope or whether exponential decrease is observed ($\kappa \to \infty$). Because data may be sparse in the tail regions, histograms can show gaps that may indicate irregular or nonmonotonic tail shape, excessive discreteness, or inadequate sampling of low-probability regions.

Estimates of tail slope are easy to obtain. A nonparametric form due to Hill[26] computes the mean of log excedences over a threshold x_0:

$$\mathrm{SLOPE} = \left\{\frac{1}{r}\sum_{j=n-r+1}^{n} \log[X_{(n-r+j)}/X_{(n-r)}]\right\}^{-1} + 1 , \tag{24}$$

obtained from the largest $r + 1$ data values, where $X_{(n-r)}$ is the largest sampled value not exceeding x_0. A simple alternative is the slope estimator in a linear regression of $\log[f(x)]$ versus $\log(x)$, using frequencies obtained from a log-log histogram. Another measure,[27] and the one computed in MCNP4A, is the maximum likelihood estimate of a generalized Pareto distribution (gPd) slope parameter κ. The generalized Pareto density function is

$$f_{gP}(x;\lambda,\kappa) = \begin{cases} \lambda^{-1}\{1 + x/[\lambda(\kappa - 1)]\}^{-\kappa} & 0 \leq x < \infty, \kappa < \infty \\ \lambda^{-1}\exp(-x/\lambda) & 0 \leq x < \infty, \kappa \to \infty \end{cases} , \tag{25}$$

where λ is a scale parameter and κ is a slope parameter, with $\kappa \to \infty$ corresponding to a limiting exponential distribution. The generalized Pareto distribution has the property that excedences over a threshold also follow a gPd, with κ unchanged.[28] The maximum likelihood slope estimator used in MCNP4A is based on the largest 200 histories, with at least three distinct values required and a maximum value for κ of 10. This procedure gives estimates roughly equivalent to those obtained by explicitly selecting a threshold. To give an ~10% relative error in the slope estimator when $\kappa = 3$, i.e., when two moments exist, 200 points were chosen (see Ref. 27 for details). Both the nonparametric and Pareto estimators are sensitive to the choice of threshold value and to the fraction of data sampled; thresholds need to be chosen sufficiently far out in the tail so that the power law is a reasonable approximation, but not so far out that data are too sparse. Furthermore, the gPd estimator appears to be more sensitive than the nonparametric estimator to clumping of tail data. In any case, whether graphical or numerical measures are used, recall that a slope <3 indicates that the theoretical variance does not exist (based on the available sampled data), and hence, the asymptotic normal-theory results for t do not apply.

II.F. Summary

Section II described some modifications to the current practice for obtaining confidence intervals: (a) use of third- or fourth-moment estimators to determine if convergence has been achieved; (b) use of skewness-corrected intervals; and (c) use of tail slope estimators and log-log density plots to indicate the number of moments that appear finite. Section III reports simulations conducted to determine the effectiveness of these procedures and to search for sample-based quantities that indicate when n

is sufficiently large for t to have converged to a standard normal.

III. SIMULATION STUDY

This section reports results from a simulation study of interval estimation with highly skewed analytic data. The first goal of the simulation was to study the relationship between coverage rate and the diagnostic statistics described in Sec. II for finite samples. The second goal was to evaluate sample-based indicators of coverage rate validity, i.e., indicators of whether n is sufficiently large that sampling can cease and valid confidence intervals for μ can be formed. Simulation distributions, described in Sec. III.A, were based on analytic Monte Carlo distributions that have behaviors typically found in particle transport problems. Both continuous and discrete distributions were used, with tail slope values chosen to obtain a variety of coverage rates.

III.A. Simulation Distributions

Pederson[17] considered the absolute value of a Cauchy (avC) random variable, truncated at a point $U > 0$. This density function is a close approximation to those found in some difficult observed Monte Carlo tallies.[11] The density has x^{-2} (i.e., slope of 2) tail behavior, and without truncation, no moments exist, but with truncation, all do. The density is

$$f_{avC}(x;U) = \frac{2}{\pi}\frac{1}{1+x^2} \quad , \quad 0 \le x < U \quad , \tag{26}$$

and

$$P(X = U) = H(U) = 1 - \frac{2}{\pi}\arctan(U) \quad , \tag{27}$$

where the remaining mass accumulates at U. Another candidate distribution is the generalized Pareto [Eq. (25)]. This distribution allows for specification of any number of finite moments, with the k'th moment existing if $\kappa > k + 1$.

To mimic a distribution that has Cauchy-like behavior for $x < U$ but then drops off at a faster rate, we use a composite absolute value Cauchy-Pareto distribution ($avCP$)

$$f_{avCP}(x;\kappa,U) = \begin{cases} f_{avC}(x;U) & 0 \le x < U \\ H(U)f_{gP}(x-U;\lambda_U,\kappa) & x \ge U \quad , \end{cases} \tag{28}$$

where the value of the scale parameter λ is chosen so that Cauchy and Pareto contributions to the density are equal at the transition point U. Note from Eq. (25) that $f_{gP}(0;\lambda,\kappa) = \lambda^{-1}$; thus, $\lambda_U = H(U)/f_{avC}(U;U)$. The density f_{avCP} is a valid density with $\int f_{avCP}(x)\,dx = 1$. As with the generalized Pareto, the k'th moment exists if the slope $\kappa > k + 1$. A suitably large value of U results in a distribution that for small sample sizes appears to have no moments existing (as with the truncated Cauchy). As n increases, however, the slope estimate should increase in value and give evidence of finite moments as the Pareto component of the tail is sampled more heavily. Figure 3 shows the log-log density plots of generalized Pareto and absolute value Cauchy-Pareto (with $U = 1000$) random variables, with tail slopes of 3.5 and 5.17, respectively. Note that eventually the Cauchy-Pareto drops off at a faster rate than the pure Pareto.

An example arising from particle transport is an analytic score distribution for a spatially continuous tridirectional Monte Carlo transport problem using the exponential transform, as described by Booth.[29] The exponential transform can be used with either analog or implicit capture. This distribution has support on a countable set of points on the positive reals $\mathcal{R}^+$, as well as probability at zero. We will focus on the nonzero values for these data. Particles are constrained to scatter either forward or backward or up or down. Isotropic scattering, typically used here, parameterizes this as follows: Particles have a $\frac{1}{4}$ probability of continuing in the same direction, a $\frac{1}{4}$ probability of reversing direction, and a $\frac{1}{2}$ probability of scattering at a right angle. Distributions are indexed by an exponential transform parameter p and a macroscopic scattering cross-section parameter σ_s. The total macroscopic cross section σ_t is always set to 1, so that the ratio $\sigma_s/\sigma_t = \sigma_s$ is the scattering probability. Figure 4 illustrates this distribution for parameters $p = 0.5$, $\sigma_s = 0.5$, isotropic scattering, and analog capture. A log-log histogram is superimposed on the plot, and the dots correspond to theoretical score probabilities for the various numbers of forward and backward collisions a particle can make. This histogram approximates a density for these scores, the tail shape of which is nearly linear, with slope of ~5.5. This lends some theoretical support for using continuous power-law distributions such as the Cauchy and Pareto to model tail behavior for actual particle transport problems. However, note that sufficiently fine binning of the histogram (not shown) reveals the discrete, sawtooth pattern of the actual score probabilities. Also, note that choosing implicit capture of particles results in a sampled distribution that, while still discrete, is much smoother than the corresponding analog case.

III.B. Simulation Design

Five different simulation problems are presented here:

1. generalized Pareto with (slope) $\kappa = 3.5$, $\lambda = 1$, two moments existing
2. absolute value Cauchy-Pareto with $U = 500$, $\kappa = 5.17$, four moments existing
3. absolute value Cauchy-Pareto with $U = 1000$, $\kappa = 5.17$, four moments existing

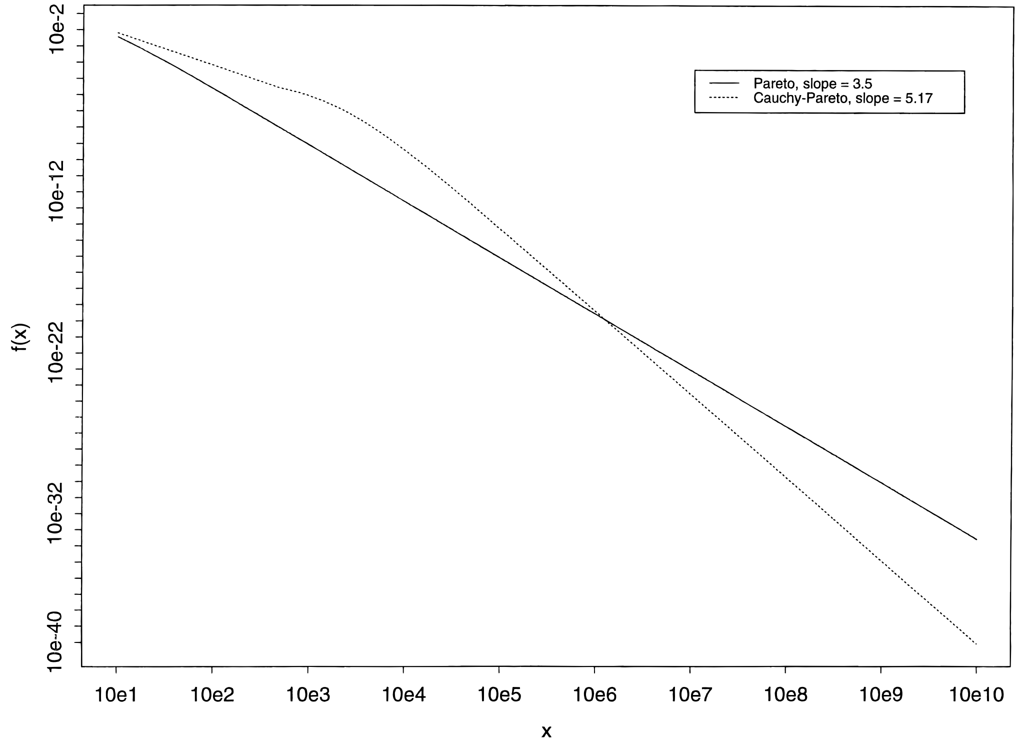

Fig. 3. Log-log density plot for Pareto and Cauchy-Pareto distributions.

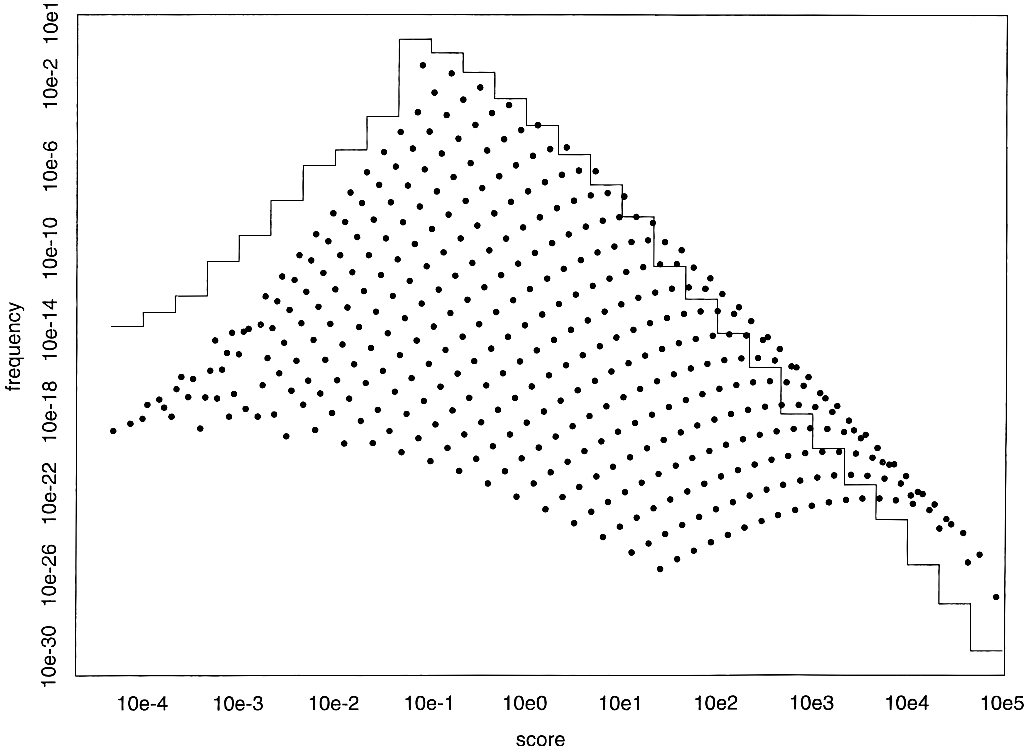

Fig. 4. Score probabilities (dots) and log-log histogram for analog tridirectional distribution, $p = 0.5$, $\sigma_s = 0.5$.

4. tridirectional spatially continuous distribution, $p = 0.5$, $\sigma_s = 0.5$, with isotropic scattering, analog capture, and four moments existing
5. tridirectional spatially continuous distribution, $p = 0.9$, $\sigma_s = 0.5$, with isotropic scattering, analog capture, and two moments existing.

Problems 1 and 5 have only two moments existing, while the remaining three problems have four finite moments, so these are all highly skewed distributions. However, we might expect problem 1, the pure generalized Pareto, to converge more quickly than problems 2 and 3 because of the Cauchy tail component for these latter problems. Problem 2 should converge more quickly than problem 3 as U is smaller. Problem 5 should be the slowest to converge, as its distribution has a very shallow tail slope and is highly discrete. Convergence rates can be obtained from the Edgeworth expansions of Sec. II.B, which require the existence of at least three moments and a continuous sampled distribution; neither condition is satisfied for problem 5. The distribution in problem 4 is also discrete, but it is smoother than that in problem 5 (and has four moments existing).

The aforementioned five problems were chosen to obtain a range of convergence behaviors for distributions with finite variance, with more interest in slow-to-converge problems. Other problems were run by the authors but are not included here; see Ref. 30 for analyses. Two of these bear mention. A truncated absolute value Cauchy with a large truncation point generated data that behaved for the most part like that from a pure absolute value Cauchy. All moment estimators, including the mean, exhibited jumps indicative of infinite moments, until sample sizes were large enough to reflect the tail truncation. A second additional distribution used was the tridirectional distribution with implicit capture and with other settings the same as in problem 5. For this problem, two moments are finite, like the analog capture case, but the additional smoothness of the distribution gives results that are qualitatively similar to those seen in problem 1 (the pure Pareto with two finite moments).

The simulation was designed as follows: For each of the five problems, two independent ranges of sample sizes n are generated. These ranges were chosen so that one range has low coverage and the other has near-nominal coverage. (This was accomplished except in problems 4 and 5; problem 4 has consistently good coverage rates, and problem 5 has consistently low coverage rates.) Four hundred independent replications were obtained for each sample size range.

Within each range, data were collected in a manner similar to that found in the MCNP tally fluctation charts described in Sec. I. At each of 20 equally spaced numbers of histories n, the following statistics were stored: n, $\bar{x}$, RE, VOV, SLOPE, and cumulative maximum sampled x_i. For problems 1, 2, and 3, the Pareto slope estimator [Eq. (25)] was used, and for problems 4 and 5, the nonparametric estimator [Eq. (24)] was used; as mentioned in Sec. II.D, the nonparametric estimator appears to be less sensitive to the clumping of tail values in highly discrete sampled distributions. The primary measure of coverage rate performance will be the observed coverage rate (OCR), the sample fraction of intervals $\mathcal{I}_u = (-\infty, u_\alpha)$ that cover the true mean μ. Nominal 95% upper-tail intervals will be used here. Results for upper 90, 97.5, and 99.5% are qualitatively similar and discussed more fully in Ref. 31. With 400 replications, an OCR with 95% true coverage has a standard error of 0.011. We will define near-nominal coverage to be within 3 standard errors of the nominal rate, so that for 95% nominal coverage, near-nominal refers to ~91.7% or higher coverage. For right-skewed problems, lower-tail confidence intervals cover at near-nominal rates and will not be analyzed here; see Ref. 31 for details. Also computed at each n is the observed modified coverage rate (OMCR) for upper-tail intervals $\mathcal{I}_u^* = (-\infty, u_\alpha^*)$ [using Eq. (23)] and the average ratio of the half-widths of the two intervals $(u_\alpha^* - \bar{x})/(u_\alpha - \bar{x})$.

III.C. Analysis of Summarized Data

Tables II through VI display the sample statistics described previously, averaged over 400 replications, for each of the problems and the ranges of n that were simulated. Average values of n are displayed for problems 4 and 5 because only the nonzero histories were used here. In each case the smallest, largest, and five intermediate sample sizes are given (20 were actually stored). Pederson[31] reproduces the full tables, along with empirical cdf's and standard errors of the various statistics. Discrepancies between low- and high-range results that have equal n's are all within acceptable sampling error limits, as would be expected.

Examination of the OCRs indicate that convergence was achieved most quickly in problem 4; both ranges have coverages near 95%. Problems 1, 2, and 3 (ranked in the order of convergence) also have near-nominal coverage for the largest n's. For each of these problems, the low ranges have coverages falling somewhat below 95%. Problem 5 does the worst, even with the largest absolute sample sizes, with the OCR not moving much above 85% for any simulated level of n. Several comments can be made about the significance of these results before we start examining their relation to RE, VOV, etc. The Cauchy part of the absolute value Cauchy-Pareto distributions in problems 2 and 3 induces considerably slower convergence than the pure Pareto distribution from problem 1 exhibits, even though the tail of the *avCP*'s eventually drops more sharply. Problem 4 is well behaved.

In problem 5, the second moment is barely finite, and the convergence to normality is extremely slow. Additional runs of this problem provided evidence that convergence is eventually achieved, albeit very slowly. For n's of 100 to 150 million, average VOV dropped to ~0.10,

TABLE II

Averages and Observed Coverage Rates, $\alpha = 0.05$, for Problem 1—Generalized Pareto, $\kappa = 3.5$, $\mu = 1.6667$

n	$\bar{x}$	RE	VOV	FOM	SLOPE	OCR	OMCR	$\frac{u^*_{0.95} - \bar{x}}{u_{0.95} - \bar{x}}$
Low Range, 400 Replications								
200	1.6673	0.1226	0.202	0.394	4.93	0.868	0.932	1.65
400	1.6719	0.0921	0.166	0.347	7.37	0.900	0.938	1.50
800	1.6744	0.0680	0.137	0.312	7.23	0.920	0.935	1.42
1 000	1.6704	0.0608	0.123	0.308	6.23	0.905	0.930	1.38
2 000	1.6664	0.0436	0.094	0.292	7.03	0.908	0.932	1.28
3 000	1.6685	0.0363	0.088	0.278	5.29	0.902	0.918	1.25
4 000	1.6677	0.0317	0.080	0.271	6.33	0.910	0.930	1.23
High Range, 400 Replications								
4 000	1.6641	0.0316	0.077	0.274	6.42	0.908	0.932	1.22
8 000	1.6662	0.0230	0.067	0.256	4.09	0.920	0.930	1.22
16 000	1.6674	0.0166	0.057	0.245	4.03	0.948	0.958	1.18
20 000	1.6676	0.0149	0.054	0.242	3.89	0.945	0.952	1.17
40 000	1.6680	0.0106	0.042	0.235	3.79	0.938	0.948	1.12
60 000	1.6667	0.0088	0.038	0.231	3.70	0.942	0.948	1.11
80 000	1.6667	0.0076	0.038	0.228	3.67	0.950	0.955	1.11

TABLE III

Averages and Observed Coverage Rates, $\alpha = 0.05$, for Problem 2—Absolute Value Cauchy-Pareto, $U = 500$, $\kappa = 5.17$, $\mu = 5.431$

n	$\bar{x}$	RE	VOV	FOM	SLOPE	OCR	OMCR	$\frac{u^*_{0.95} - \bar{x}}{u_{0.95} - \bar{x}}$
Low Range, 400 Replications								
800	5.412	0.239	0.443	0.0364	2.56	0.698	0.808	2.92
1 600	5.478	0.198	0.392	0.0233	2.28	0.778	0.875	2.73
3 200	5.418	0.149	0.297	0.0175	2.21	0.822	0.900	2.11
4 000	5.385	0.136	0.273	0.0163	2.20	0.840	0.915	1.99
8 000	5.433	0.104	0.200	0.0136	2.26	0.858	0.930	1.62
12 000	5.444	0.088	0.169	0.0125	2.30	0.885	0.922	1.49
16 000	5.449	0.077	0.148	0.0119	2.34	0.900	0.932	1.44
High Range, 400 Replications								
8 000	5.427	0.105	0.210	0.0134	2.25	0.852	0.922	1.67
16 000	5.434	0.077	0.147	0.0119	2.33	0.908	0.935	1.42
32 000	5.429	0.055	0.096	0.0111	2.51	0.908	0.940	1.25
40 000	5.434	0.050	0.084	0.0108	2.59	0.922	0.948	1.23
80 000	5.434	0.036	0.056	0.0103	3.00	0.922	0.938	1.16
120 000	5.435	0.029	0.043	0.0101	3.33	0.925	0.938	1.13
160 000	5.427	0.025	0.035	0.0101	3.70	0.920	0.942	1.11

TABLE IV

Averages and Observed Coverage Rates, $\alpha = 0.05$, for Problem 3—Absolute Value Cauchy-Pareto, $U = 1000$, $\kappa = 5.17$, $\mu = 5.832$

n	$\bar{x}$	RE	VOV	FOM	SLOPE	OCR	OMCR	$\frac{u^*_{0.95} - \bar{x}}{u_{0.95} - \bar{x}}$
Low Range, 400 Replications								
800	5.760	0.261	0.487	0.0344	2.77	0.630	0.795	2.97
1 600	5.723	0.216	0.435	0.0217	2.31	0.692	0.822	2.91
3 200	5.773	0.175	0.368	0.0143	2.15	0.760	0.862	2.57
4 000	5.835	0.167	0.354	0.0122	2.11	0.805	0.900	2.47
8 000	5.842	0.129	0.279	0.0097	2.16	0.815	0.900	1.99
12 000	5.838	0.110	0.239	0.0086	2.16	0.845	0.905	1.81
16 000	5.859	0.097	0.209	0.0079	2.17	0.855	0.925	1.68
High Range, 400 Replications								
8 000	5.838	0.128	0.278	0.0094	2.14	0.838	0.920	1.96
16 000	5.852	0.097	0.210	0.0078	2.17	0.872	0.915	1.67
32 000	5.853	0.071	0.156	0.0069	2.23	0.888	0.940	1.47
40 000	5.852	0.064	0.138	0.0067	2.25	0.900	0.945	1.39
80 000	5.872	0.046	0.095	0.0062	2.35	0.915	0.952	1.29
120 000	5.874	0.038	0.073	0.0061	2.45	0.932	0.948	1.22
160 000	5.875	0.033	0.062	0.0060	2.57	0.920	0.950	1.20

TABLE V

Averages and Observed Coverage Rates, $\alpha = 0.05$, for Problem 4—Tridirectional Scattering, Analog Capture, $p = 0.5$, $\mu = 0.12685$

n	$\bar{x}$	RE	VOV	FOM	SLOPE	OCR	OMCR	$\frac{u^*_{0.95} - \bar{x}}{u_{0.95} - \bar{x}}$
Low Range, 400 Replications								
262	0.1268	0.0433	0.1076	2.29	4.82	0.900	0.920	1.27
526	0.1267	0.0313	0.0788	2.11	4.10	0.908	0.945	1.19
1 055	0.1269	0.0225	0.0582	1.99	4.47	0.935	0.945	1.19
1 318	0.1269	0.0202	0.0526	1.97	4.61	0.915	0.932	1.18
2 637	0.1271	0.0144	0.0351	1.91	4.95	0.925	0.940	1.10
3 955	0.1269	0.0117	0.0260	1.90	5.21	0.952	0.955	1.08
5 271	0.1270	0.0101	0.0201	1.89	5.60	0.952	0.955	1.06
High Range, 400 Replications								
5 269	0.1269	0.0100	0.0159	1.92	5.51	0.955	0.958	1.06
10 543	0.1269	0.0071	0.0107	1.90	5.19	0.955	0.965	1.04
21 067	0.1269	0.0050	0.0064	1.88	5.71	0.958	0.965	1.03
26 328	0.1269	0.0045	0.0056	1.88	5.74	0.962	0.962	1.03
52 646	0.1269	0.0032	0.0030	1.88	5.75	0.958	0.965	1.02
78 985	0.1269	0.0026	0.0024	1.87	5.74	0.970	0.970	1.02
105 305	0.1269	0.0023	0.0018	1.87	5.75	0.962	0.965	1.01

TABLE VI

Averages and Observed Coverage Rates, $\alpha = 0.05$, for Problem 5—Tridirectional Scattering, Analog Capture, $p = 0.9$, $\mu = 0.024971$

n	$\bar{x}$	RE	VOV	FOM	SLOPE	OCR	OMCR	$\dfrac{u^*_{0.95} - \bar{x}}{u_{0.95} - \bar{x}}$
Low Range, 400 Replications								
26 736	0.02493	0.0405	0.220	0.0385	2.67	0.830	0.900	1.55
53 477	0.02499	0.0317	0.187	0.0322	2.32	0.835	0.895	1.59
106 963	0.02498	0.0249	0.216	0.0271	2.82	0.818	0.850	1.83
133 694	0.02500	0.0229	0.232	0.0250	2.90	0.845	0.865	1.94
267 408	0.02501	0.0180	0.233	0.0209	3.02	0.828	0.860	2.04
401 106	0.02500	0.0153	0.224	0.0182	3.11	0.842	0.872	1.58
534 799	0.02498	0.0132	0.199	0.0175	3.19	0.828	0.858	1.45
High Range, 400 Replications								
133 715	0.02496	0.0223	0.220	0.0260	2.94	0.798	0.818	1.89
267 403	0.02496	0.0173	0.233	0.0215	3.08	0.800	0.830	1.99
534 781	0.02499	0.0142	0.217	0.0176	3.22	0.845	0.875	1.50
668 454	0.02503	0.0140	0.212	0.0163	3.29	0.832	0.860	1.50
1 336 896	0.02499	0.0100	0.172	0.0141	3.53	0.845	0.892	1.49
2 005 392	0.02498	0.0083	0.170	0.0129	3.64	0.838	0.872	1.55
2 673 908	0.02498	0.0073	0.173	0.0121	3.71	0.842	0.890	1.65

and OCRs edged up to ~88%. Some periodicity, with amplitude decreasing as n increases, was seen in these results, similar to that evident in Table VI. This is the result of the discrete sawtooth nature of the sampled distribution, as successively more of the "teeth" are discovered by sampling. Thus, the discreteness inherent in $f(x)$ for problem 5, combined with a barely finite second moment, strongly affects the coverage rates. In problem 4, where at least four moments exist, a similar phenomenon is not observed.

In all problems, the skewness-modified confidence intervals improve coverage rates, and in problems 1 through 4, they result in near-nominal coverage for sample sizes n about half as large as those needed for nominal coverage of standard intervals. Likewise, for a given n, the amount of error in the coverage rate was roughly halved by using the modified intervals. The penalty for improved coverage is wider intervals, however; the final columns of Tables II through VI show average ratios of half-widths $(u^*_{0.95} - \bar{x})/(u_{0.95} - \bar{x})$. For problems 1, 2, and 3, modified confidence intervals range from nearly three times as large for the smallest n's to 10 to 20% larger, at which point the standard intervals cover at nearly 95%. Problem 5 has modified intervals 1.5 to 2 times as wide as standard intervals, and in fact average width goes up as n increases for certain values of n. Problem 4 has average ratios near 1 for n's with coverage near 95%. Problems 1 and 5 do not have a finite third moment, and the sampled distribution in problem 5 is highly discrete; in these cases, the Edgeworth expansion used to obtain the modified intervals is not valid, and the coverage rates for these intervals may be conservative as n grows large. Coverages do seem reasonable for problem 1, which has only two moments existing but has a continuous underlying distribution.

Based on results from Sec. II, average VOV should be a good problem-independent indicator of nominal coverage. This holds for all problems. In general, when the average VOV drops to roughly the 0.05 to 0.07 range, coverage is within 2 to 3% of 0.95. The converse also appears true; in no case when coverage was under 90%, say, was average VOV < 0.10. Figure 5 is a plot of average VOV versus OCR for the midpoint and final n's for each of the five problems and two ranges simulated. High and low ranges are denoted by h and l; for example, problem 5, high range, is represented by $5h$. For distributions with infinite fourth moment (problems 1 and 5), these results hold but should certainly be treated with caution, as the sample average VOV may not converge smoothly to zero, if at all. Also note that the rate at which average VOV drops is faster [and should be $\sim O(n^{-1})$] for the cases where convergence has been achieved and four moments exist. Decreases slower than $O(n^{-1})$ indicate a negative bias (i.e., values are smaller than expected). The third-moment analog to VOV, $\hat{\zeta} n^{-1/2}$, exhibits a similar relationship with coverage rates; see Ref. 31 for details.

Average RE is not a good indicator of coverage-rate validity, even though RE and VOV are highly correlated.

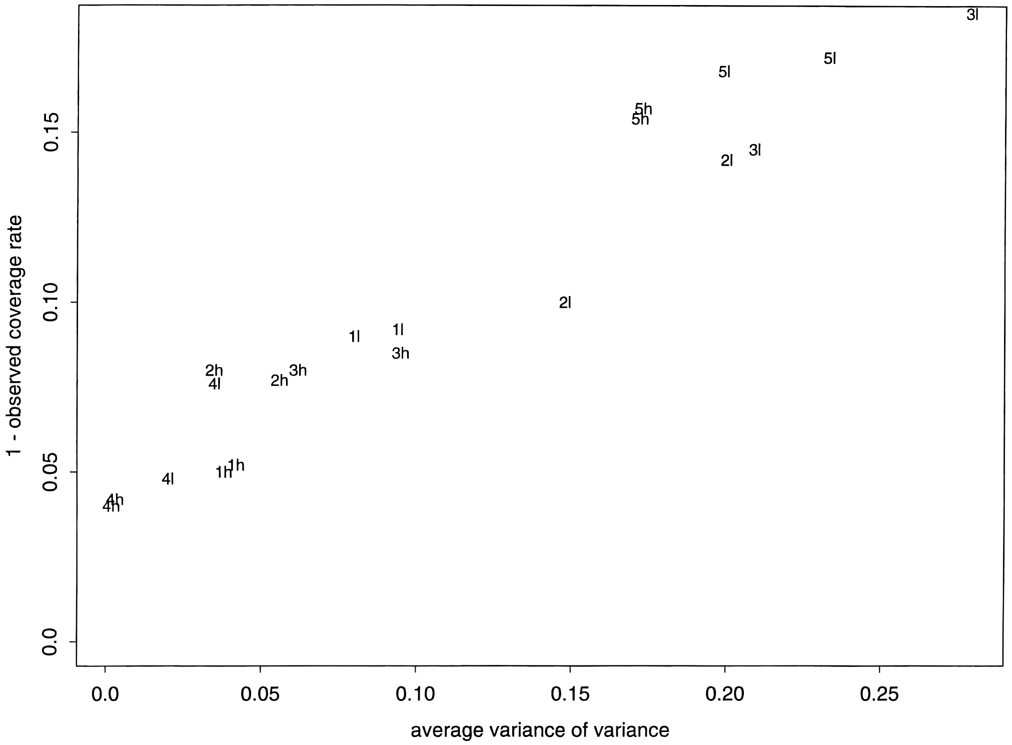

Fig. 5. Observed coverage rate versus average VOV, nominal coverage = 0.05.

For problem 5, average REs are <0.01, yet OCRs (and OMCRs) remain below 90%. The rate at which RE decreases should be $O(n^{-1/2})$ if two moments exist and convergence of the second-moment estimator has occurred. As with VOV, a slower rate of decrease indicates a negative bias in RE.

Conversely, large REs (say, >0.10) generally indicate that sampling should continue. The relative error of the mean is useful for scaling the variability in $\bar{x}$, i.e., indicating the approximate range of errors. In problems 1 and 5, average REs are <0.01 for large n; even if confidence intervals are not covering μ often enough, as in problem 5, the total error is no more than a few percent. This follows by observing that the relative biases in $\bar{x}$ are much smaller in magnitude than the relative errors observed, even for noncoverged situations. For example, for problem 5 with $n = 160\,000$, the observed relative bias in $\bar{x}$ is 0.0007, compared with an average relative error of 0.025. (Observed biases in $\bar{x}$ can be obtained in Tables II through VI by subtracting the theoretical value of μ from the average value of $\bar{x}$; relative biases are obtained by dividing this total by $\bar{x}$.)

Average FOMs appear to converge down to a fixed value for the near-nominal-coverage cases. This indicates a negative small-sample bias in s^2; $\bar{x}$ has less variance and converges more quickly than s^2, hence most of the variation in FOM is due to s^2. This is more empirical evidence for the argument that in highly skewed situations, sample variance behavior (in particular, its variation as opposed to its magnitude) is what determines the rate at which confidence intervals cover at nominal levels.

The Pareto-based slope estimator was used in the continuous data problems (1, 2, and 3) and the nonparametric estimator in problems 4 and 5. For these latter problems, the Pareto estimator was too sensitive to the choice of a truncation value because of the small number of distinct x values. Average SLOPEs generally converge to their true values, except for problems 2 and 3, where the extreme Pareto tail was insufficiently sampled. Except for problem 1, for smaller n's the bias in SLOPE is negative because of inadequate tail sampling, giving the indication that fewer moments exist than is actually the case. It should be emphasized that in general the slope estimate will be less biased than moment estimates, especially higher moments.

Summarizing, these results provide further support for using VOV (or some other measure of the variability of s^2) to indicate when confidence interval convergence has been reached. Slope estimates are useful in characterizing whether or not the moment condition of the convergence theory has been satisfied. The modified intervals were also found to bring significant improvements in coverage, with often substantial increases in interval width. The level of discreteness present in the data was seen to have a detrimental effect on coverage rate validity when only two moments exist. We next examine batched-means normality tests for one of the simulated problems, followed by a discussion of the use of individual sample quantities as indicators of coverage rate validity.

III.C.1. Batched-Means Normality Tests

Consider simulation problem 1 (generalized Pareto distribution, two moments existing). Along with the statistics given in Table II, we computed the modified Wilk-Shapiro test W' from Eq. (9) for each sample, based on $m = 25$, 50, and 100 batches. Results are given in Table VII [columns are sample size n, average VOV and OCR, average W' for each of three numbers of batches $m = 100$, 50, and 25, and rejection rates of W' for these numbers of batches (with corresponding batch sizes B)]. The tests were at $\alpha = 0.05$, for the hypothesis that the batch means were generated from a Gaussian distribution. Critical values of the test were (0.975, 0.953, 0.918), corresponding to $m = (100, 50, 25)$. Critical values for the test for other values of m can be easily generated and are available from the first author (S.P.P.).

Several specific observations are immediately apparent. The likelihood of rejecting the normality hypothesis is directly related to the number of batches m that were used. For example, consider the three cases where $B = 160$. When $m = 100$ batches were generated, this test rejected normality 83% of the time, while when 25 batches were used, only one-third of the cases resulted in rejection (even though a batch sample size of 160 clearly does not correspond to a converged situation). This highlights a characteristic of all statistical goodness-of-fit tests; if the sample size is large enough, any departure from the assumed distribution (Gaussian, in this case) will result in rejection. Hence, if the batch size B is not sufficiently large and is fixed as m grows, the hypothesis of normality may never be accepted because increasing m will detect finer and finer discrepancies from normality.

A second comment relates to the intended use of the test, which is for assessing normality of the overall mean $\bar{x}$. The Wilk-Shapiro test rejection rate shows considerable dependence on both m and B: For $m = 100$, most cases reject normality, while for $m = 25$, <50% do. Both these facts are true regardless of whether the overall mean $\bar{x}$ has converged ($n > 10\,000$) or not ($n < 10\,000$). The modified Wilk-Shapiro test W' is not indicating when $\bar{x}$ has converged. Instead, the test accomplishes what it is designed to do: indicate when the distribution of a random variable is normal. However, here this indication is for the batched means based on samples of (batch) size B and not for the overall mean with sample size n. For $n = 80\,000$ and $m = 25$, batch sizes are $B = 3200$, and convergence to normality is beginning to occur (20% of the samples reject normality, when nominally 5% should). Sampling would need to continue much longer (in this case, a factor of 5 to 10 times as long) using the stringent Wilk-Shapiro statistic than our empirically based criteria indicate. Note that our criteria do not necessarily indicate full convergence to normality, but rather only that approximately valid confidence intervals for $\bar{x}$ can be formed.

Similar phenomena were seen for the other distributions simulated in this paper. The distributions considered here are in some sense "best case" because they are the most regular and exhibit smooth convergence to normality. Normality tests using batched means depend on the number of batches considered and provide limited information about whether the full mean $\bar{x}$ has converged, at least in the sense of covering the true value at nominal levels. W' did not consistently indicate normality in converged situations. There is still a place for batched-mean–normality tests in assessing convergence, however. For this problem, it appears that when $m = 50$ there is some direct correspondence between OCR and the W' rejection rate; however, in other models different levels of m gave better results. Statistics like W' can be used in a similar manner to our use of VOV and the slope estimator in obtaining rules about stopping strategies. Q-q plots are a useful graphical tool in assessing convergence but again, for means of size B and not necessarily of size n. In short, normality tests can be useful when combined

TABLE VII

Batched-Mean Wilk-Shapiro Test Averages and Rejection Rates, $\alpha = 0.05$, for Problem 1—Generalized Pareto, 400 Replications, $\kappa = 3.5$, $\mu = 1.6667$, m = Number of Batches

			Average W'			Rejection Rate (and Batch Size B)		
n	VOV	OCR	$m = 100$	$m = 50$	$m = 25$	$m = 100$	$m = 50$	$m = 25$
1 000	0.123	0.905	0.789	0.850	0.881	1.000 (10)	0.888 (40)	0.463 (40)
4 000	0.080	0.910	0.862	0.904	0.911	0.970 (40)	0.555 (80)	0.333 (160)
8 000	0.067	0.920	0.892	0.930	0.934	0.938 (80)	0.445 (160)	0.278 (320)
16 000	0.057	0.948	0.909	0.935	0.934	0.830 (160)	0.513 (320)	0.255 (640)
40 000	0.042	0.938	0.935	0.941	0.948	0.775 (400)	0.280 (800)	0.303 (1600)
80 000	0.038	0.950	0.945	0.960	0.947	0.680 (800)	0.280 (1600)	0.205 (3200)

with other tools, but by themselves they are poor indicators of when confidence intervals based on $\bar{x}$ cover at nominal levels and may be very inefficient when used as stopping criteria.

III.D. Analysis of Run-Specific Quantities

The second main goal of the simulation was to ascertain if sample-based quantities can predict when coverage rates will be near nominal levels. As Efron[32] points out, in one sense this is an impossible problem because there can always be a small amount of probability, sufficiently far out in the tail of the distribution, to cause sample means to underestimate μ. However, it is also true that if the central limit theorem conditions are satisfied so that t has an $\mathcal{N}(0,1)$ distribution, eventually intervals will cover at the nominal rate; this analysis is merely an attempt to predict when convergence has occurred. In particular, we wish to obtain estimates of $P(\mu \in \mathcal{I}_\alpha)$, for $\mathcal{I}_\alpha = (-\infty, u_\alpha)$, where u_α is the upper end point of a $(1 - \alpha) \times 100\%$ upper confidence interval for μ as defined earlier. In this formula, as in all confidence interval formulas, the random quantity is not μ but $\mathcal{I}_\alpha$.

Care must be taken when defining "validity" with respect to confidence intervals. An individual sample produces an interval that either covers μ or does not; rather, we are interested in predicting whether or not the probability of coverage is near the nominal value $1 - \alpha$. Furthermore, this probability should be conditioned on the value of any predictors that are chosen, e.g., $P(\mu \in \mathcal{I}_\alpha | \text{VOV} < 0.05) = 1 - \alpha$. In fact, conditioning should not be based directly on sample size n because we are effectively trying to determine if a given n is large enough.

How conditioning is done is important for the following reason. Consider the rule alluded to earlier: Stop sampling if $\text{VOV} < 0.05$, and continue running the problem otherwise. From the results presented so far, this seems like a reasonable rule. However, for a given sample size n it may be true that the larger VOVs correspond to "better" coverage than do smaller VOVs (the same holds true for REs). For example, in problem 2, with $n = 160\,000$, overall coverage of μ is 0.92, fairly close to nominal. Of the samples run (out of 400) with $\text{VOV} < 0.05$, 91% of the resulting intervals cover, while if $\text{VOV} > 0.05$, 97% cover. This makes sense because a smaller VOV is correlated with a smaller RE and hence smaller intervals. Nevertheless, this result is only useful for characterizing coverage for this specific value of n and for the specific distribution of this problem. Across problems and sample sizes, smaller VOVs are indicative of coverages closer to nominal values. In searching for rules that are problem- and sample size-independent, we will instead look at both "converged" and "nonconverged" situations and compare the behavior of predictors in each case.

Prediction rules of the following form will be considered: Stop sampling if some condition A is satisfied, and continue sampling otherwise. These rules can be made up of other rules; we will focus mainly on AND-type rules; i.e., A is an intersection of other conditions. In particle transport simulations, stopping too soon with an incorrect value of a parameter is generally more costly than running a simulation for too long. Thus, we will prefer rules that suggest stopping only when there is strong evidence that n is sufficiently large. The set of possible predictors examined here will be taken from a set of statistical checks similar to those that appear in the current implementation of MCNP.

III.D.1. Statistical Checks

The following list from Forster[19] is a set of checks or rules based on information contained in TFCs in MCNP4A. In each case, the checks are passed if the stated condition is satisfied. When not specified, the checks refer to the value of statistics at the end of the run.

1. $\bar{x}$ was not monotonically increasing or decreasing over the last half of the simulation run.

2. $\text{RE} <$ threshold (usually 0.10 or 0.05).

3. RE decreases at a rate of $\sim n^{-1/2}$ over the last half of the run.

4. RE is monotonically decreasing, with small fluctuations allowed, over the last half of the run.

5. $\text{VOV} <$ threshold (usually 0.10).

6. VOV decreases at a rate of $\sim n^{-1}$ over the last half of the run.

7. RE is monotonically decreasing, with small fluctuations allowed, over the last half of the run.

8. $\Delta\text{FOM} <$ threshold (usually 0.20), where ΔFOM is defined as the relative range of 1/FOM over the last half of the run; i.e., $\Delta\text{FOM} = [\max(\text{FOM}^{-0.5}) - \min(\text{FOM}^{-0.5})]/(\text{FOM}^{-0.5}$ at termination). This parameterization was chosen to accentuate the variation in FOM due to variation in s^2.

9. FOM was not monotonically increasing or decreasing over the last half of the simulation run.

10. $\text{SLOPE} > 3$ (at least two moments are indicated to exist).

Checks 2, 5, and 10 are calculated at the end of a run; the remaining checks incorporate intermediate-n information. All of these checks, many based on heuristic evidence built up over time, are designed to pass when the $\mathcal{N}(0,1)$ approximation to the distribution of t is valid. As $\bar{x}$ nears μ, it should fluctuate randomly. RE should drop at rate $n^{-1/2}$ when $s/\bar{x}$ is close to σ/μ, and dropping nearly monotonically is equivalent to the fluctations in $s/\bar{x}$ being small relative to the decrease in $n^{-1/2}$. Similar comments apply to the checks involving VOV, the relevant quantity in this case being $\hat{\mu}_4/s^4$, which converges to μ_4/σ^4. The FOM rules monitor the stability and randomness of $s/\bar{x}$. Analysis of these and other rules follows.

III.D.2. Analysis of Checks

For the purpose of developing predictors, we will assume that convergence has been achieved for the high ranges of problems 1, 2, and 3 and both ranges of problem 4, based on OCRs at the end of each of those simulations. Similarly, we will consider the other five problem-range combinations to be nonconverged. Within these broad groupings, we can order the combinations by the closeness to 95% coverage: 4-high, 1-high, 4-low, 2-high, 3-high for the converged problems; and 1-low, 2-low, 3-low, 5-high, 5-low for the nonconverged problems. (These groupings are completely arbitrary. In particular, note that 1-low and 3-high have very similar OCRs. Different groupings may result in different rules being developed, but the qualitative content will remain similar. Likewise, rules developed for OMCRs instead of OCRs will give different thresholds but will be of similar nature. In Sec. IV, we include some rules based on OMCRs instead of OCRs.) Rules should thus indicate stopping for the converged problems with high probability and stopping for the nonconverged problems with low probability. We will first consider single-condition rules and then move on to combinations (pairs, triples, etc.).

Of the ten checks described previously, the VOV threshold check is the single best predictor of convergence, when evaluated purely in terms of observed coverage rate. Two other measures, ΔFOM and the rate of decrease in RE, convey virtually the same information as each other and are also valuable predictors. (The relevant quantity for the rate of decrease in RE is not the absolute decrease but rather the distance from the expected decrease rate of $1/\sqrt{2} = 0.7071$ for a doubling of n.) Figures 6, 7, and 8 are plots of empirical cdf's of these three statistics for the ten problem-range combinations, based on 400 replications. (For example, "1*l*" stands for problem 1, low range.) Ideal predictors will have empirical cdf values of 1 for converged situations and empirical cdf values of 0 for nonconverged cases. The classification and regression tree (or CART) approach of Brieman et al.[33] was used to determine decision rules for these predictors. CART verified that checks 5, 8, and 3 (with check 5, the VOV < threshold rule, most powerful) are the best single-check predictors in terms of having the lowest misclassification rates. Separating lines, or thresholds, for these predictors from CART are 0.045, 0.105, and 0.060, respectively, and are included as vertical lines on the plots. Based on these data, it appears that separating lines somewhere between 0.03 and 0.05 for VOV, between 0.08 and 0.12 for ΔFOM, and between

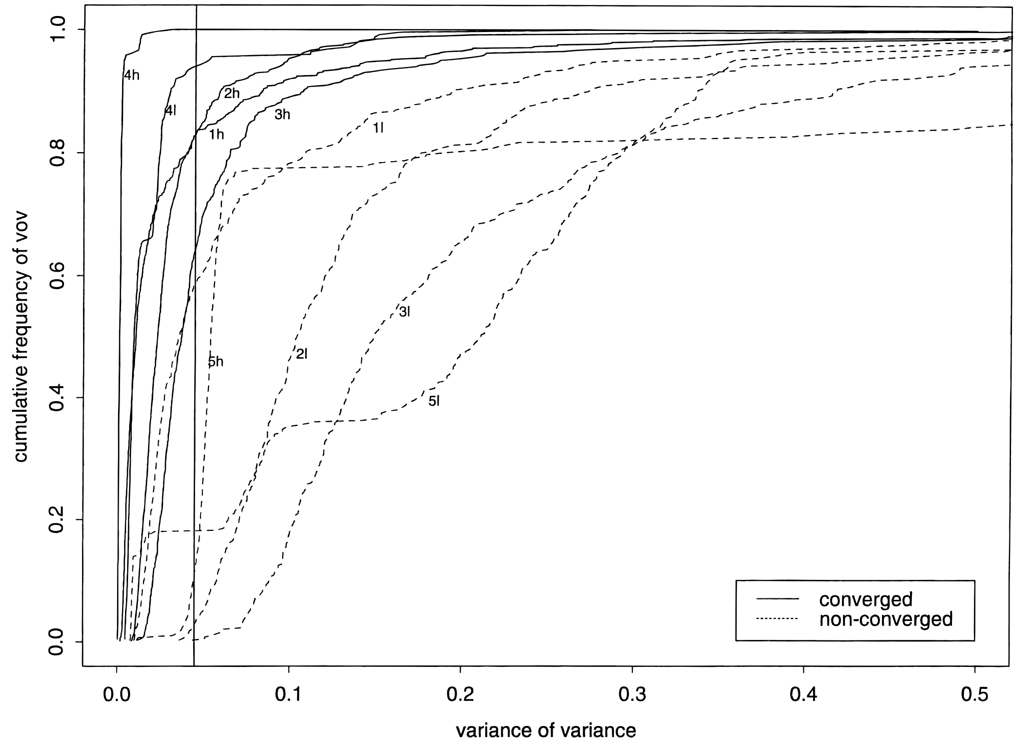

Fig. 6. Empirical cdf's of VOV, with vertical separating line.

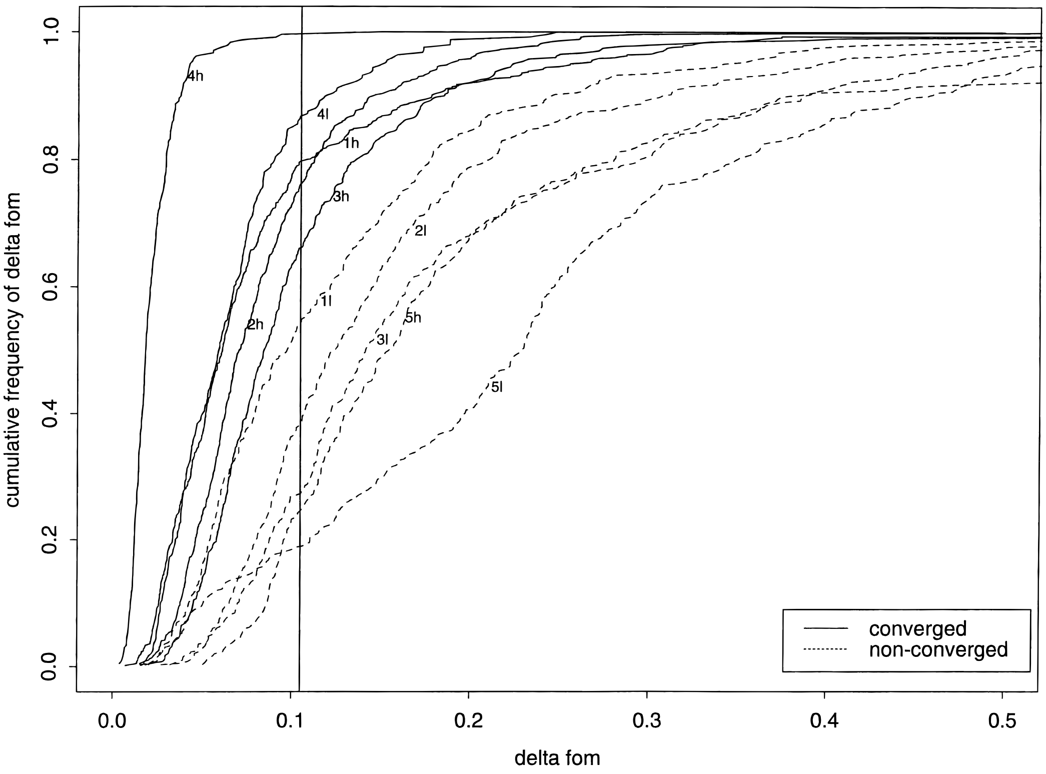

Fig. 7. Empirical cdf's of ΔFOM, with vertical separating line.

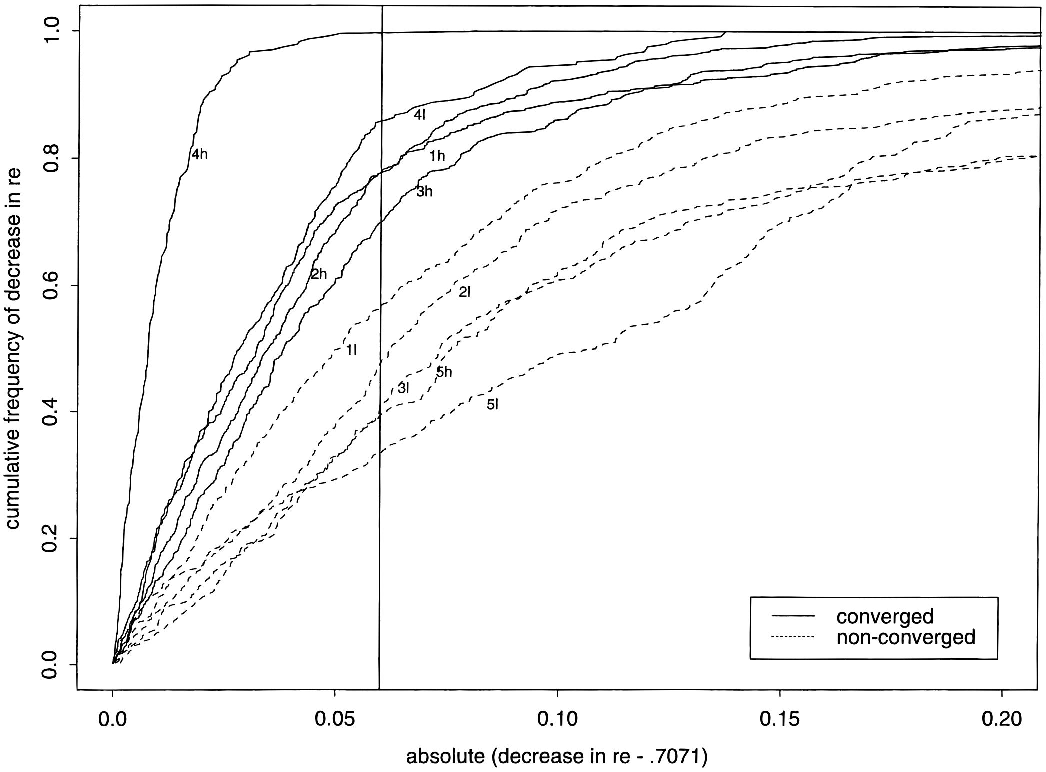

Fig. 8. Empirical cdf's of decrease in RE, with vertical separating line.

0.05 and 0.08 for the decrease in RE are appropriate. No other single check from the aforementioned list of ten had strong predictive power, although some of the remaining checks will be discussed later. (See Ref. 31 for more details on this analysis.) For what follows, we will focus on ΔFOM and note that groupings based on the decrease in RE were virtually identical to those based on ΔFOM.

Testing pairs of rules with the AND construct again found the combination of VOV and ΔFOM to be most significant in predicting coverage rate validity. Table VIII shows individual pass/no pass rates for each problem, for VOV < 0.045, for ΔFOM < 0.105, and for the rule that combines those two rules: VOV < 0.045 and ΔFOM < 0.105. (Fine tuning of these joint thresholds may result in slightly higher classification rates, but this may result in overfitting of the available data and is not recommended for generalization purposes.) For this combined rule, the pass rates for the ten groups give an ordering very similar to that given at the start of this subsection. The only potential problem is with the 17% pass rate for the low range of problem 5. These runs corresponded to very low values of VOV and ΔFOM, with SLOPE indicating several moments existing, a situation for which these rules would suggest that sampling should cease. These are cases where the tail of the density has not been adequately sampled. Lowering the VOV threshold would filter these cases out but will bring down overall pass rates.

No triple or larger set of checks improved the separation of groups considerably. The decrease in RE gave the same classifications as did ΔFOM and did not interact differently with the VOV rule. Checks 1, 3, 6, and 8 had no substantial predictive power, and neither did check 2, the RE threshold rule.

There was limited evidence that the rate of decrease in VOV (check 7) was closer to the expected rate of $1/n$ for converged samples than for nonconverged samples when, in addition, the slope estimator (check 10) indicated that at least two moments exist (SLOPE > 3). However, the low range of problem 5 was not correctly classified using this rule. Neither check 7 nor check 10 were useful single indicators of convergence, but each provides information about the number of finite moments, particularly the slope estimator.

Summarizing, the most useful measures found here for prediction of confidence interval validity were effectively surrogates for the variability in the sample variance s^2. VOV measures this directly, and the ΔFOM and rate of decrease in RE measure the stability of $s/\bar{x}$, which measures the stability of s because of the faster convergence of $\bar{x}$. Scale measures (such as RE) and measures based on convergence patterns were not found to be useful predictors.

We have found rules that use VOV and work even when there is strong evidence that fewer than four moments exist. The properties of estimators like VOV (effectively, a scaled-moment estimator divided by sample size: $\hat{\mu}_4/s^2n$) in these cases is in general not known, but rates of convergence depend on the degree of smoothness of the underlying distribution. When the underlying moment (in this case, μ_4) of an estimator such as VOV does not exist, there is no guarantee that large jumps will not occur in the value of the estimator, or at what rate they may occur. Caution should be taken if VOV is to be used in these settings. An extreme case is problem 5, where severe skewness combines with discreteness to bias the behavior of both moment and slope estimates. This bias is most pronounced for functions of higher order moments like VOV. For this problem, VOV gets "fooled" in the low-n range, while ΔFOM does not (note that the second moment does exist). Discreteness can also bias the true tail slope but to a lesser degree, as the next maxima to be sampled may be an order of magnitude larger than anything seen previously, with a corresponding order of magnitude drop in probability. From the point at which diagnostic statistics satisfy the aforementioned rules, it is recommended that the simulation

TABLE VIII

Pass Rates for Selected Rules

Problem	VOV < 0.045	ΔFOM < 0.105	VOV < 0.045 and ΔFOM < 0.105
5 low	0.180	0.188	0.172
5 high	0.072	0.338	0.040
3 low	0.002	0.272	0.002
2 low	0.025	0.382	0.022
1 low	0.585	0.542	0.478
3 high	0.640	0.660	0.555
2 high	0.828	0.762	0.712
4 low	0.940	0.868	0.848
1 high	0.828	0.798	0.782
4 high	1.000	0.995	0.995

be run (or examined, if the runs have already been made) a further period of time (1.5 or 2 times the original n) and SLOPE, VOV, and other statistics that measure variability in s^2 be monitored. If these diagnostics still look reasonable after this further calculation, there is additional evidence that the problem has truly converged.

All of the predictors in this section are random quantities and hence subject to sampling bias and statistical error. Higher order sample moments are more likely to be biased than lower ones. Even in what appear to be converged settings, large jumps in the value of statistics such as VOV or RE can occur if the corresponding population moments do not exist. These jumps can be based on a single extreme observation. Jumps can also occur when higher order moments exist but become less likely as n increases. Any sort of inferential procedure will only be based on the available data, and if sufficient unsampled $f(x)$ tail probability exists, eventually it will be discovered, sometimes in the form of a single very large observation. This indicates that even when conditions suggest that sampling can cease, and with four or more moments indicated to exist, additional particle histories should be generated past the point that these conditions were first satisfied. Running the problem longer will reduce the bias in the statistics, if relevant moments exist. If relevant moments do not exist, additional sampling will still reduce the chance that the observed value of the statistic is too low. For example, examination of the high range of problem 5 in Table VI suggests that if we consider those samples that passed the VOV < 0.045 rule at average $n = 534\,781$ (of which there are 14%), only 25% of these samples also passed when n was doubled (see Ref. 31 for additional discussion). For the low range of problem 1, which is on the edge of being a converged problem, a similar technique applied at $n = 4000$ (with a pass rate for the combined rule of 48%) found that 90% of these samples passed at the higher n of 8000. Note that for both of these sample sizes, coverage with the modified interval is very close to the nominal 95%, so the high pass rate for a situation termed "nonconverged" is not a problem.

Some other sample-based measures, such as the ratio of the cumulative maximum value to the cumulative sum and various functions of the slope estimator, were also tried as predictors, with limited success. Both measure the influence of large values with low probability, and more research is needed in both areas. We close by noting that other techniques commonly used in simulation to improve convergence properties, such as the use of variances of batched means to estimate σ^2, were tried but had no effect on overall coverage rates.

IV. CONCLUSIONS

IV.A. Example Revisited

We now return to the Monte Carlo test problem discussed in Sec. I. The problem required much larger samples before inference stabilized. The tally fluctuation chart for the entire 1-billion-history run is given in Table IX for two different point detector tallies (tallies 35 and 45) that estimate the same quantity and are independent of each other. Our discussion will concern tally 35, the one we originally discussed, but similar comments apply to tally 45. After ~400 million histories, the VOV < 0.045 and ΔFOM < 0.105 rules are satisfied, and the (standard) 90% confidence interval for μ at this point, using Eq. (7), is $5.774 \times 10^{-8} \pm 0.256 \times 10^{-8}$ (or 5.517×10^{-8}, 6.030×10^{-8}), which includes the true μ of 5.64×10^{-8}. However, the SLOPE at this point (2.2) indicates only the mean existing, so sampling should continue; the RE may still experience large jumps. Eventually, at ~650 million histories, SLOPE exceeds 3, indicating a finite variance, with the other rules still satisfied. Running the problem to 1 billion points (not quite the doubling suggested in Sec. III.D) again satisfies all rules and gives more indication that four moments exist. In retrospect, the user could probably have stopped at 650 million histories, after an indication that more than two moments exist, but with the proviso that the VOV estimator is subject to possible large jumps. However, running the calculation longer at this point would still be prudent. Note that we have intentionally chosen an extreme case to test our recommendations; in most Monte Carlo particle transport tallies, the variance generally exists, and four or more moments often do. (For this problem, superior estimators are available. In fact, the two-dimensional ring detector is 230 times more efficient than the point detector, and the one-dimensional surface flux estimator is 28 500 times so.) The methodology presented here is general and is designed to be applied in any situation with iid histories.

IV.B. Recommendations

The theoretical and empirical results presented here are intended to provide guidance for the Monte Carlo practitioner in making inferences about the simulated problem. It is assumed that the user will run the calculation for a period of time. At that point, the following recommendations regarding the formation of valid confidence intervals can be applied:

1. Plot the estimated tally density $f(x)$. Log-log histograms and tail density slope estimates, such as those produced in MCNP, can provide visual information regarding the number of moments that appear to exist, the appropriateness of the power-law approximation for the tail, and the degree of discreteness present in the data.

2. Monitor the VOV, the relative change in the FOM, the rate of decrease in the RE, and SLOPE. With VOV $<$ 0.03 to 0.05, ΔFOM $<$ 0.08 to 0.12, and an indication of four or more moments existing (SLOPE > 5), the results of Sec. III indicate a high likelihood that the problem in question has converged. Alternatively, replacing the

TABLE IX

Large-n TFCs for Example Problem—Point Detector Tally, $\mu = 5.64 \times 10^{-8}(\pm 0.02\%)$

n	$\bar{x}$	RE	VOV	FOM	SLOPE
Tally 35					
65 536 000	6.4201E-08[a]	0.0673	0.0897	1.3	2.0
131 072 000	5.8210E-08	0.0427	0.0588	1.6	2.3
196 608 000	5.8982E-08	0.0373	0.0733	1.4	2.1
262 144 000	5.6879E-08	0.0306	0.0615	1.6	2.2
327 680 000	5.7017E-08	0.0278	0.0455	1.5	2.2
393 216 000	5.7736E-08	0.0270	0.0347	1.3	2.2
458 752 000	5.7618E-08	0.0245	0.0297	1.4	2.3
524 288 000	5.7136E-08	0.0231	0.0256	1.4	2.4
589 824 000	5.7067E-08	0.0215	0.0221	1.4	2.7
655 360 000	5.6790E-08	0.0202	0.0199	1.4	3.1
720 896 000	5.7927E-08	0.0200	0.0167	1.3	4.0
786 432 000	5.7358E-08	0.0186	0.0161	1.4	3.8
851 968 000	5.7515E-08	0.0184	0.0161	1.3	4.0
917 504 000	5.7470E-08	0.0176	0.0148	1.3	4.2
983 040 000	5.7540E-08	0.0173	0.0147	1.3	4.9
1 000 000 000	5.7745E-08	0.0172	0.0141	1.3	5.4
Tally 45					
65 536 000	5.4740E-08	0.0496	0.1167	2.4	2.6
131 072 000	5.4582E-08	0.0399	0.0670	1.9	2.6
196 608 000	5.3922E-08	0.0334	0.0612	1.8	2.2
262 144 000	5.5989E-08	0.0318	0.0530	1.5	1.9
327 680 000	5.6103E-08	0.0288	0.0456	1.4	1.8
393 216 000	5.5560E-08	0.0256	0.0388	1.5	2.0
458 752 000	5.5768E-08	0.0233	0.0316	1.6	2.4
524 288 000	5.5128E-08	0.0212	0.0285	1.6	2.3
589 824 000	5.5585E-08	0.0215	0.0313	1.4	2.4
655 360 000	5.5971E-08	0.0206	0.0259	1.4	2.6
720 896 000	5.5810E-08	0.0195	0.0234	1.4	2.9
786 432 000	5.5655E-08	0.0184	0.0215	1.4	3.4
851 968 000	5.5595E-08	0.0178	0.0196	1.4	3.8
917 504 000	5.5649E-08	0.0172	0.0178	1.4	4.9
983 040 000	5.5567E-08	0.0166	0.0165	1.4	5.2
1 000 000 000	5.5419E-08	0.0164	0.0164	1.4	5.1

[a] Read as 6.4201×10^{-8}.

aforementioned VOV rule with VOV < 0.10 and the ΔFOM rule with $\Delta\text{FOM} < 0.20$ (and keeping the SLOPE rule the same) will indicate convergence when using the modified confidence intervals. If the first two conditions are satisfied, but the slope estimator indicates fewer than four moments existing, caution should be taken when forming confidence intervals, as large jumps in the value of sample statistics can occur (see recommendation 4). If fewer than two moments are indicated to exist, sampling should definitely continue to determine if the variance is finite. In addition, if histograms indicate a high degree of discreteness or clumping in the data, higher order moment estimators (particularly RE and VOV) and the slope estimator (to a lesser degree) may be subject to bias unless there is a strong indication that at least four moments exist. Large jumps in the value of one of the higher order moments indicate that more sampling is needed for that particular moment estimator to converge.

3. Use the skewness-modified confidence interval $\mathcal{I}_u^*$ if three moments are indicated to exist. These intervals give asymmetric bounds about the sample mean, with improved coverage. In general this is a safe procedure to use (even when only two moments exist); when

convergence of t to normality has been achieved, the modified interval is about the same length as the standard interval and covers at about the same rate. When less than three moments are indicated to exist, the asymptotic result is not valid, but the resulting intervals will be conservative (and possibly subject to large increases in width). Likewise the result, which is based on an Edgeworth expansion, is not asymptotically valid for discrete sampled distributions, but in practice this appears to matter only when less than four moments are finite. In any case, the modified intervals will be at least as conservative as the standard, uncorrected intervals. (Note that for all cases considered herein, the OMCR is never <0.92 when average VOV < 0.1 as required by MCNP's check 5 in Sec. III.D.1. Thus the user should have very good confidence in the OMCR.)

4. Although it is not necessary in most cases, an added measure of confidence can always be obtained by running still more histories. When is this reasonable and how many more histories should be run? Suppose that the user has run n histories and the conditions described in recommendation 2 are first satisfied for $n' < n$. If $n < 2n'$ and fewer than four moments are indicated to exist at n histories, then extend the calculation to $2n'$ histories; if $n > 2n'$ already, then no additional histories are required. If $n < 1.5n'$ and four moments are indicated to exist at n histories, then extend the calculation to $1.5n'$ histories; if $n > 1.5n'$ already, then no additional histories are required. In either case, sample statistics should be monitored and sampling stopped only if the aforementioned criteria regarding the variability of s^2 are satisfied. Stopping at this point will not absolutely guarantee that convergence has been achieved, but it does give the user more evidence that the requirements needed for application of central limit theorem–type results are satisfied. This strategy, while perhaps conservative, will tend to identify highly biased cases, as was seen for the low-range n's for problem 5. If the conditions in recommendation 2 are not satisfied, and especially if there is no evidence of two finite moments, more histories should be examined.

5. The relative error of the mean RE should be thought of primarily as a measure of the error $\bar{x}$, as opposed to an indicator of convergence. However, confidence intervals definitely should not be formed if $RE > 0.1$.

6. Batched-mean normality tests can be used to monitor the distribution of batch means of size B. If means of size B are normally distributed, it follows that means of size $n = mB$ will be also. However, the converse does not follow, and thus batched means tests are not recommended for determining when to stop sampling.

In summary, the variability in the sample variance s^2, measured by VOV (when four moments exist) or other surrogate statistics, is the primary determinant of confidence interval validity. The rate of convergence to approximate normality of t is influenced by both the numbers of moments which are finite, measured by an estimate of the tail $f(x)$ slope, and the degree of discreteness in the underlying sampled distribution. Thus, diagnostics that measure these quantities will be useful in making inferences when sampling from highly skewed populations of the sort that commonly appear in particle transport simulations. It is the authors' hope that this paper will stimulate the use of statistical methodology among transport simulation practitioners, and that feedback from users will lead to further refinements and discoveries.

REFERENCES

1. "MCNP—A General Monte Carlo N-Particle Transport Code: Version 4A," LA-12625-M, J. F. BRIESMEISTER, Ed., Los Alamos National Laboratory (1993).

2. R. A. FORSTER, T. E. BOOTH, and S. P. PEDERSON, "A New Method to Assess Monte Carlo Convergence," *Proc. Seminar Advanced Monte Carlo Computer Programs for Radiation Transport*, Saclay, France, April 27–29, 1993, Organization for Economic Cooperation and Nuclear Energy Agency (1993).

3. R. C. GEARY, *Biometrika*, **34**, 209 (1948).

4. J. PICKANDS, *Ann. Stat.*, **3**, 119 (1975).

5. W. WILLINGER, M. S. TAQQU, W. E. LELAND, and D. V. WILSON, *Stat. Sci.*, **10**, 67 (1995).

6. R. J. SERFLING, *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York (1980).

7. P. BILLINGSLEY, *Probability and Measure*, John Wiley & Sons, New York (1979).

8. I. M. SOBOL, *Comp. Methods Math. Phys.*, **33**, 1391 (1993).

9. K. KANG and B. SCHMEISER, *Oper. Res.*, **38**, 546 (1990).

10. R. G. SARGENT, K. KANG, and D. GOLDSMAN, *Oper. Res.*, **40**, 898 (1992).

11. R. A. FORSTER, S. P. PEDERSON, and T. E. BOOTH, "Ten New Checks to Assess the Statistical Quality of Monte Carlo Solutions in MCNP," *Proc. 8th Int. Conf. Radiation Shielding*, Arlington, Texas, April 24–28, 1994, Vol. 1, p. 414, American Nuclear Society (1994).

12. I. LUX and L. KOBLINGER, *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*, CRC Press, Boca Raton, Florida (1991).

13. A. DUBI, in *CRC Handbook of Nuclear Reactors Calculations*, Vol. II, CRC Press, Boca Raton, Florida (1987).

14. A. DUBI, "On the Analysis of the Variance in Monte Carlo Calculations," *Nucl. Sci. Eng.*, **72**, 108 (1979).

15. R. D. COOK and S. WEISBERG, *Residuals and Influence in Regression*, Chapman Hall, New York (1983).

16. P. HALL, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York (1992).

17. S. P. PEDERSON, "Mean Estimation in Highly Skewed Samples," LA-12114-MS, Los Alamos National Laboratory (1991).

18. G. P. ESTES and E. D. CASHWELL, "MCNP1B Variance Error Estimation," TD-6-27-78, Los Alamos National Laboratory (1978).

19. R. A. FORSTER, "A New Method of Assessing the Statistical Convergence of Monte Carlo Solutions," *Trans. Am. Nucl. Soc.*, **64**, 305 (1991).

20. K. NOACK, *Ann. Nucl. Energy*, **64**, 305 (1991).

21. W. G. COCHRAN, *Sampling Techniques*, 3rd ed., John Wiley & Sons, New York (1977).

22. M. H. KALOS and P. A. WHITLOCK, *Monte Carlo Methods*, p. 27, John Wiley & Sons, New York (1986).

23. N. J. JOHNSON, *J. Am. Stat. Assoc.*, **73**, 536 (1978).

24. P. HALL, *Ann. Stat.*, **11**, 569 (1985).

25. P. HALL, *J. Royal Stat. Soc.*, Ser. B, **54**, 221 (1992).

26. B. M. HILL, *Ann. Stat.*, **3**, 1165 (1974).

27. J. R. M. HOSKING and J. R. WALLIS, *Technometrics*, **29**, 339 (1991).

28. W. H. DuMOUCHEL, *Ann. Stat.*, **11**, 1019 (1983).

29. T. E. BOOTH, "Analytic Score Distributions for a Spatially Continuous Tridirectional Monte Carlo Transport Problem," *Nucl. Sci. Eng.*, **122**, 79 (1996).

30. S. P. PEDERSON, "Interval Estimation in Highly Skewed Data" (1995) (unpublished manuscript).

31. S. P. PEDERSON, "Confidence Interval Results for Some Distributions Used in Monte Carlo Particle Transport" (1995) (unpublished manuscript).

32. B. EFRON, *Can. J. Stat.*, **9**, 139 (1981).

33. L. BRIEMAN, J. H. FRIEDMAN, R. A. OLSHEN, and C. A. STONE, *Classification and Regression Trees*, Wadsworth International Group, Belmont, California (1984).